Out of Sample Validation of Structured Expert Judgment

Roger M Cooke, Justin Eggstaff
Resources for the Future, Univ. Strathclyde
USMC

COST workshop
Univ Strathclyde
Aug 29, 2014

Classical Model (1991)

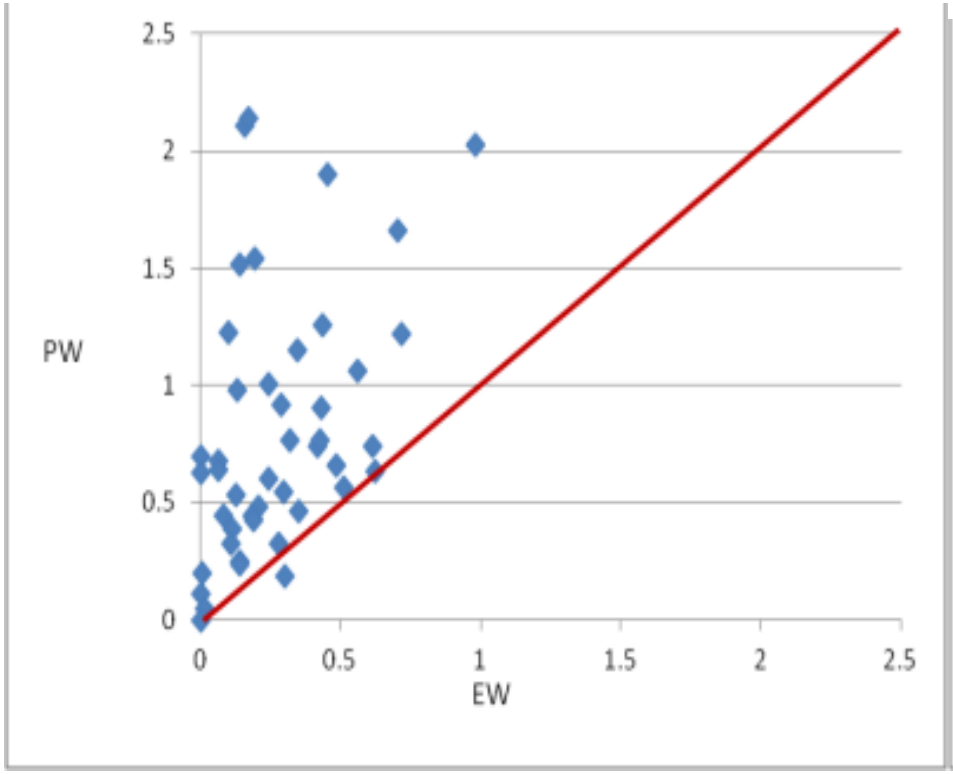
- In addition to vbls of Interest, Exprts give quantiles on seed vbls **FROM THEIR FIELD** [~10] whose values are known post hoc
- Experts scored wrt calibration (statistical accuracy) and informativeness
- Performance based weights (PW) (asymptotic strictly proper scoring rule) compared with Equal Weights (EW)
- Data sets from ~ 100 studies available

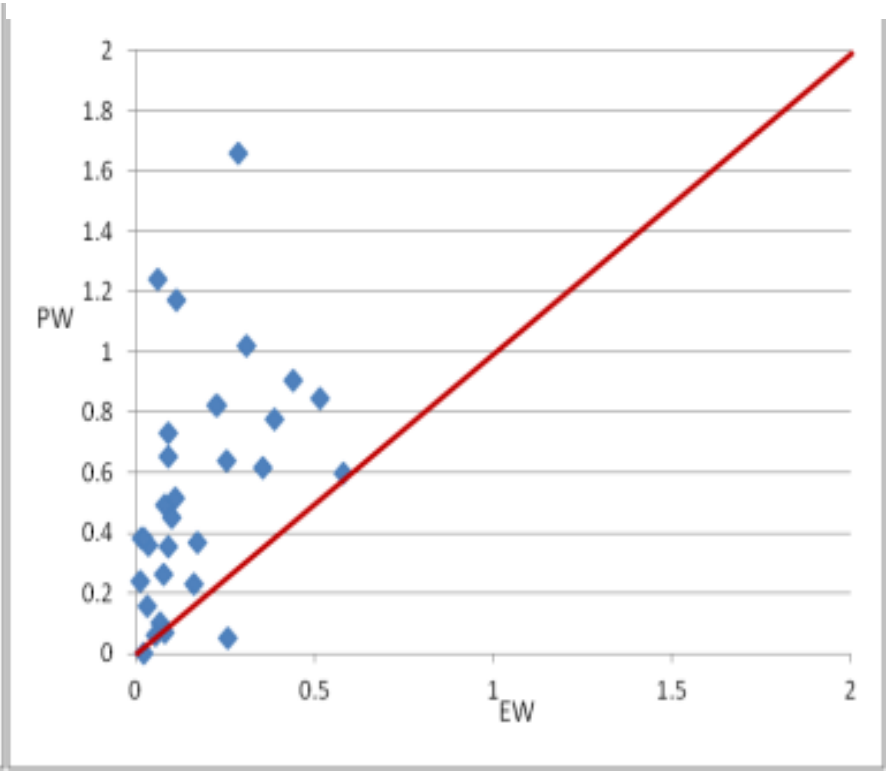
Combined score (calibration * information) Performance & Equal Weights

pre 2008

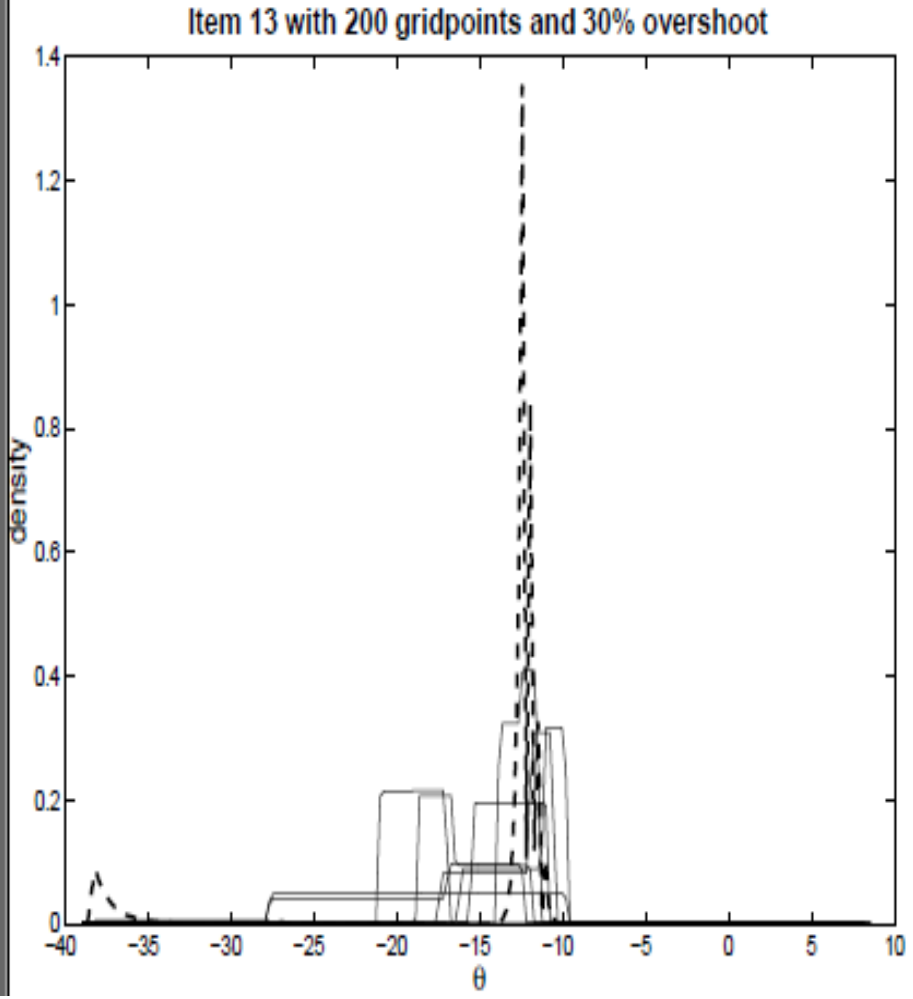
post 2008

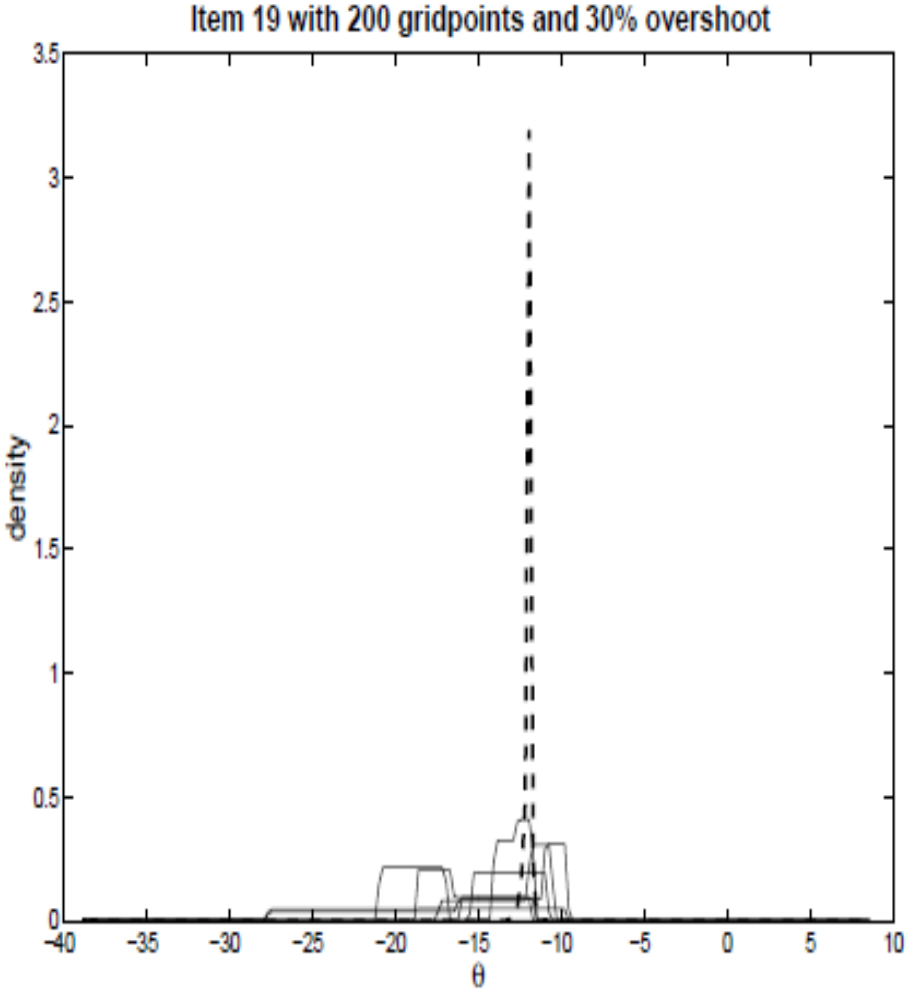
In-sample validation





In sample validation NOT foregone conclusion: Jouni and Clemen 1996





	Combined DM's:			Experts:	
	global	equal	item	best	Clemen
Calibration	0.36	0.15	0.9	0.13	0.0001
Information	1.24	0.894	1.116	1.276	2.534
Combination	0.4443	0.1341	1.005	0.1659	2.534×10^{-3}

TABLE 4: Expert Calibration results for the EUNRCDIS case.

Out of sample validation?

DO NOT use Remove-One-at-a-Time (ROAT)

- Expert 1 $P_{heads} = 0.8$ Expert 2 $P_{heads} = 0.2$
- Weights w_1/w_2 = likelihood ratio Ex1 / Ex2

N Heads & N Tails,

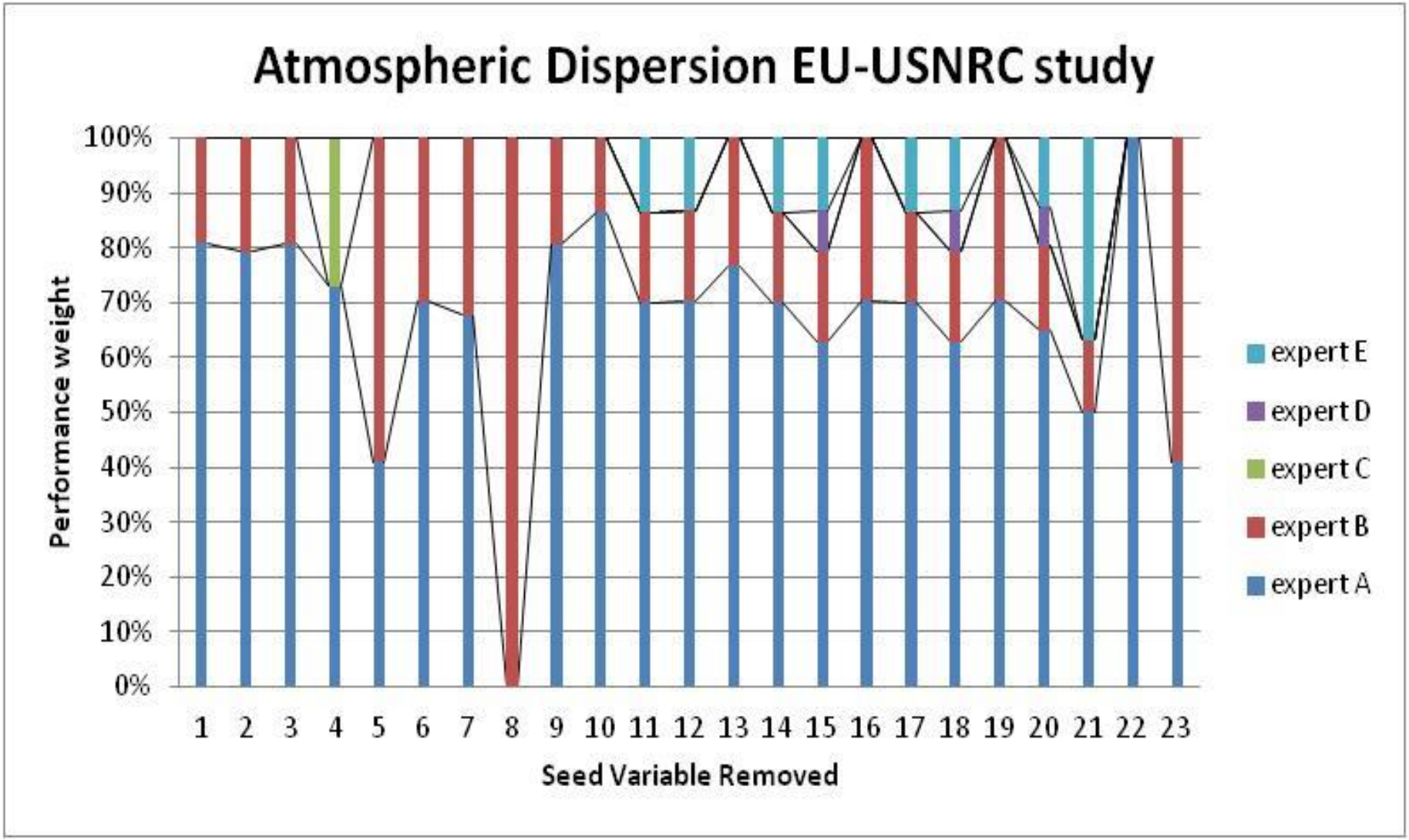
LR = $0.8^N \times 0.2^N / 0.2^N \times 0.8^N = 1.$

Remove **one H**, LR = 0.2/0.8 = ¼ = w_1/w_2

$P^{DM}{}_{heads} = (1/5) \times 0.8 + (4/5) \times 0.2 = 0.32.$

Use this to predict removed item? BIAS = $(0.32/0.5)^N$

Variation of expert weights under one-at-a-time seed variable exclusion.



ROAT Used by

- Cooke, R.M. (2008) Special issue on expert judgment, Editor's Introduction Reliability Engineering & System Safety, 93, Available online 12 March 2007, Volume 93, Issue 5, May 2008, Pages 655-656.
- Clemen, R.T (2008)" Comment on Cooke's classical method" Reliability Engineering & System Safety, Volume 93, Issue 5, May 2008, Pages 760-765
- Lin, Shi-Woei, and Bier, V.M. (2008) "A Study of Expert Overconfidence" Reliability Engineering & System Safety, 93, 775-777, Available online 12 March 2007. Volume 93, Issue 5.
- Lin, Shi-Woei, Cheng, Chih-Hsing (2008) "Can Cooke's Model Sift Out Better Experts and Produce Well-Calibrated Aggregated Probabilities?" Department of Business Administration, Yuan Ze University, Chung-Li, Taiwan Proceedings of the 2008 IEEE IEEM
- Lin, Shi-Woei, Cheng, Chih-Hsing (2009) "The reliability of aggregated probability judgments obtained through Cooke's classical model", Journal of Modelling in Management, Vol. 4 Iss: 2, pp.149 – 161,
- Shi-Woei Lin, Ssu-Wei Huang, (2012) "Effects of overconfidence and dependence on aggregated probability judgments", Journal of Modelling in Management, Vol. 7 Iss: 1, pp.6 - 22
- Cooke, R.M. (2012) "Pitfalls of ROAT Cross Validation Comment on Effects of Overconfidence and Dependence on Aggregated Probability Judgments" ,Journal of Modelling in Management, vol.7, nr. 1, pp 20-22, ISSN 1746-5664.

CROSS Validation

- Cooke, R.M., (2008) Response to Comments, Special issue on expert judgment Reliability Engineering & System Safety, 93, 775-777, Available online 12 March 2007. Volume 93, Issue 5, May 2008.
- Flandoli, F. Giorgi W.P. Aspinall, W. and Neri A (2010). " Comparing the performance of different expert elicitation models using a cross-validation technique" appearing in Reliability engineering and System Safety.
- Eggstaff,J.W., Mazzuchi,T.A. Sarkani, S. (2013) The Effect of the Number of Seed Variables on the Performance of Cooke's Classical Model, Reliability Engineering and System Safety 121 (2014) 72–82.
- Burgman, M. et al (20??) Intelligence Game, IARPA shoot-out

Out-of-sample Cross Validtion

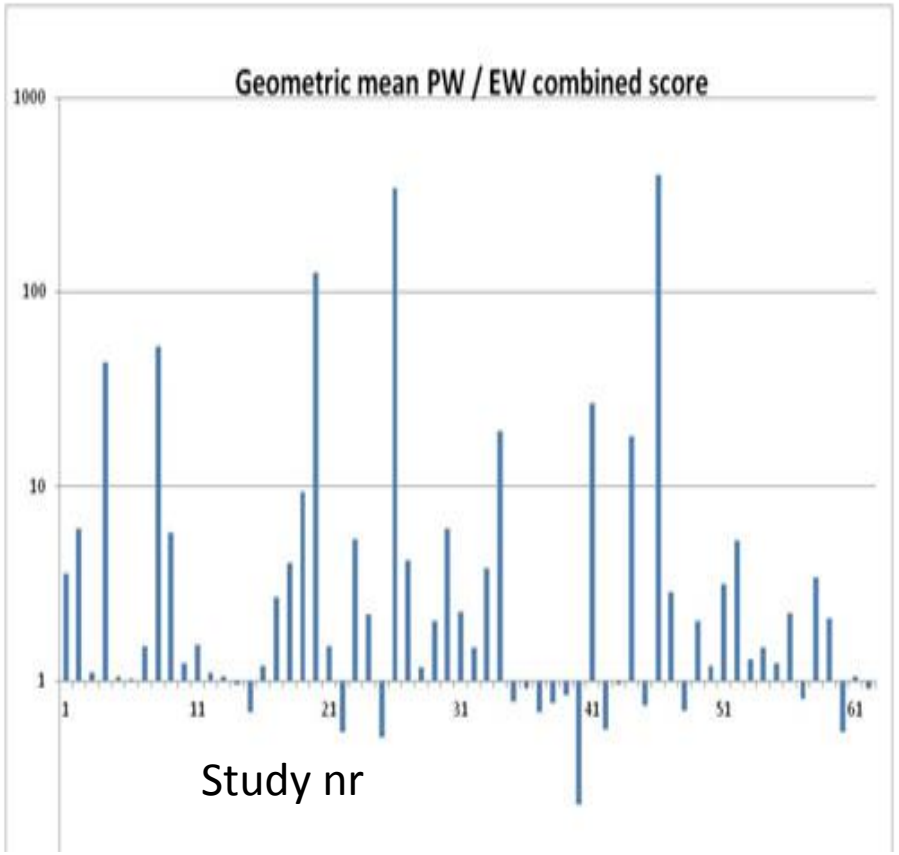
- N seed vbls
- K < N training set; N-K test set
- **WHICH K?**

- K small, low power to resolve experts
- K large, low power to resolve DM
- K = N-1, ROAT bias
- K = N/2…all k-tuples Law of Large Numbers??

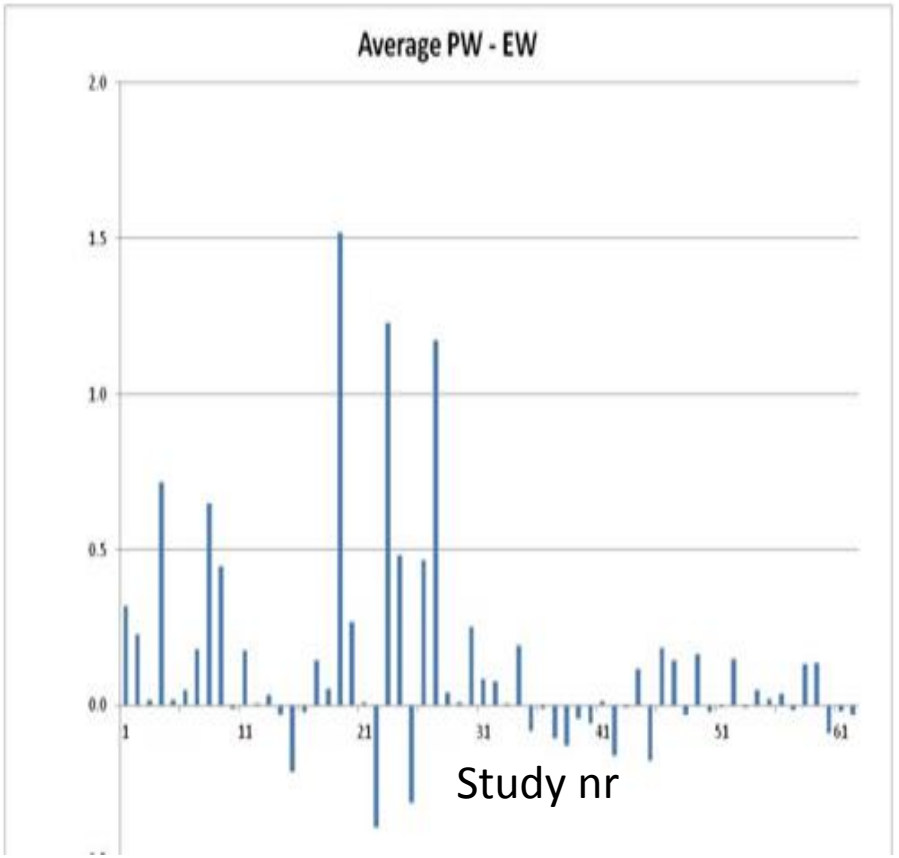
Eggstaff et al

- For K = 1….#seeds = N;
  - Initialize on EACH training sets size K
  - Score PW and EW on each test set
  - For given K average PW and EW scores
- Aggregate over all K by
  - Arithmean of PW-EW [affected by statistical power loss as K ↗]
  - Geomean of PW/EW [better, dimensionless]

%(PW > EW) = 73% (Eggstaff et al)



Smallest to largest # seeds

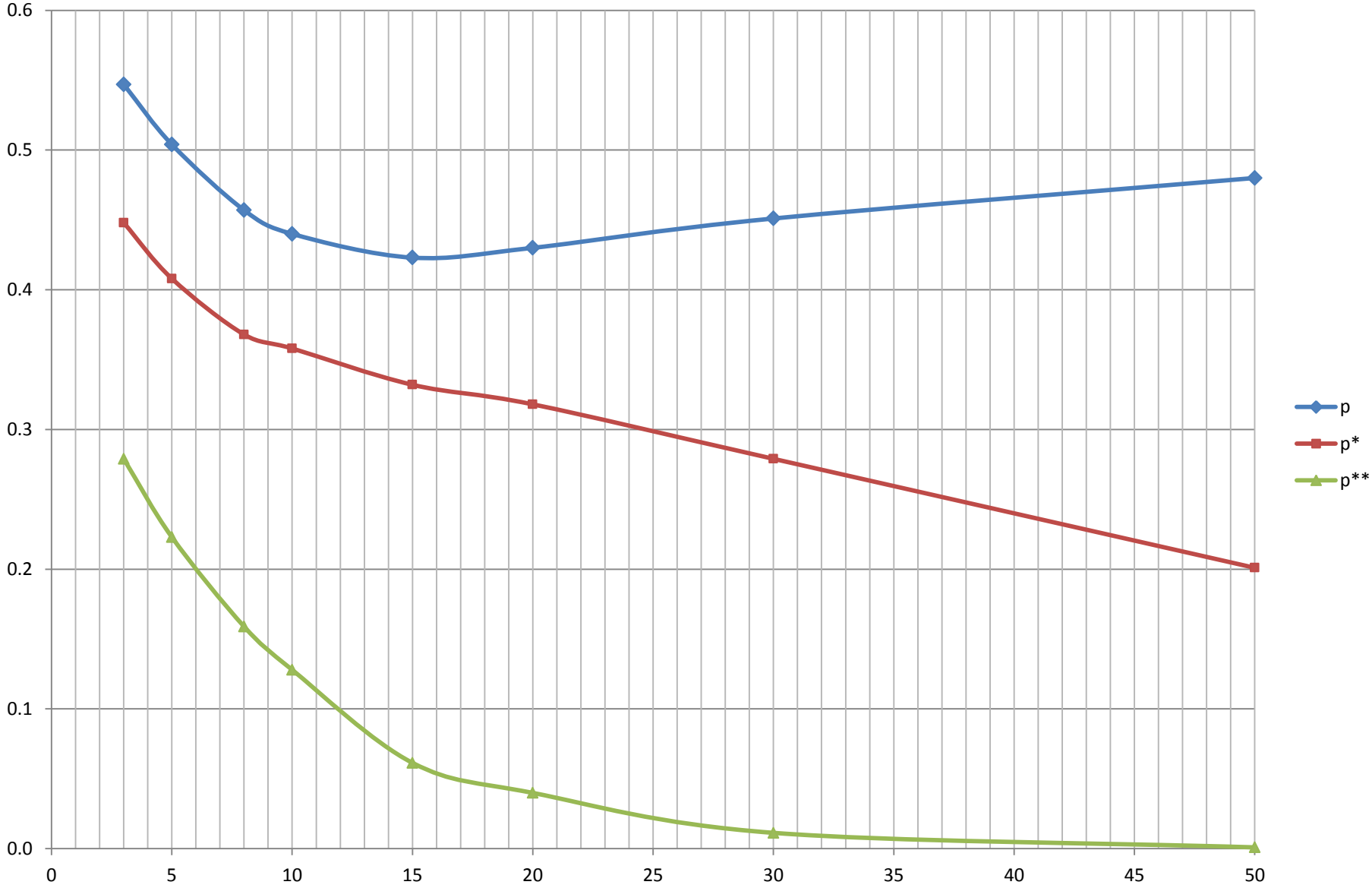


Smallest to largest # seeds

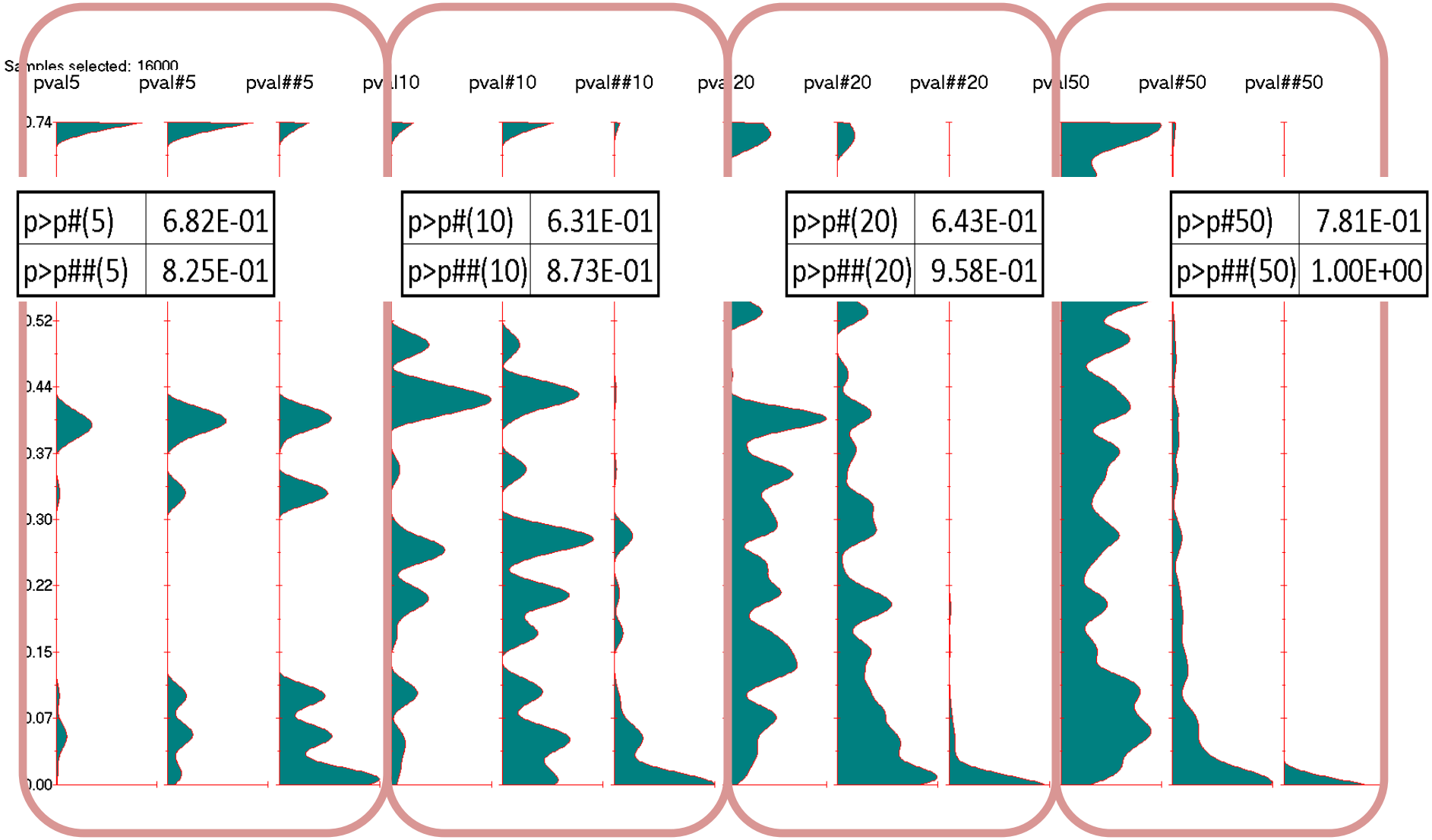
for realizations in [0.05, 0.50, 0.95] interquantile intervals

- **DM P is perfectly calibrated (5%, 45% 45%, 5%)**
- **DM P* has prob.(10%, 40%, 40%, 10%)**
- **DM P** has prob.(20%, 30%, 30%, 20%)**

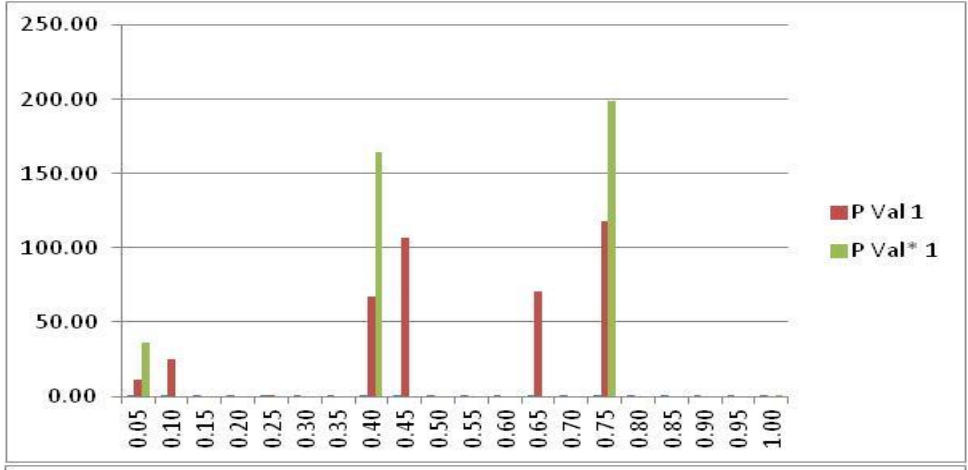
Mean p-value of P, P*, P** as function of nr seeds



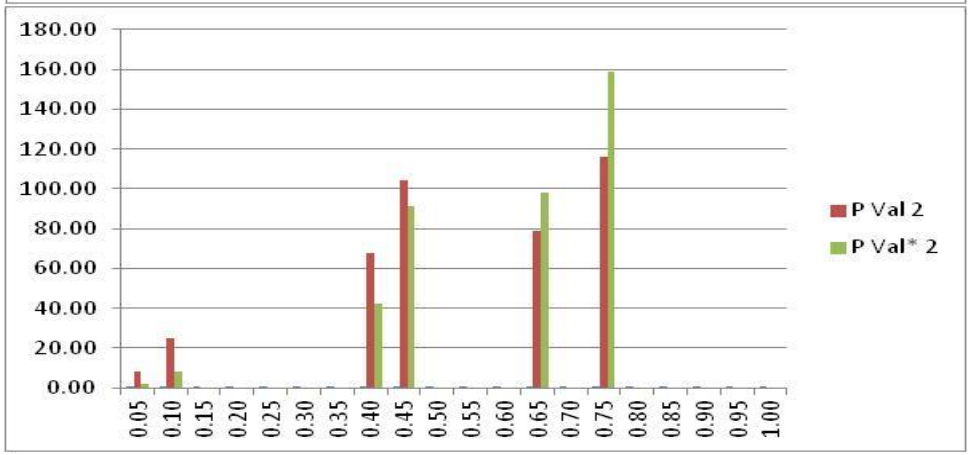
P-values with 5,10,20,50 samples
for p=(5,45,45,5); $p^{\#}$ = (10,40,40,10); $p^{\#\#}$=(20,30,30,20)



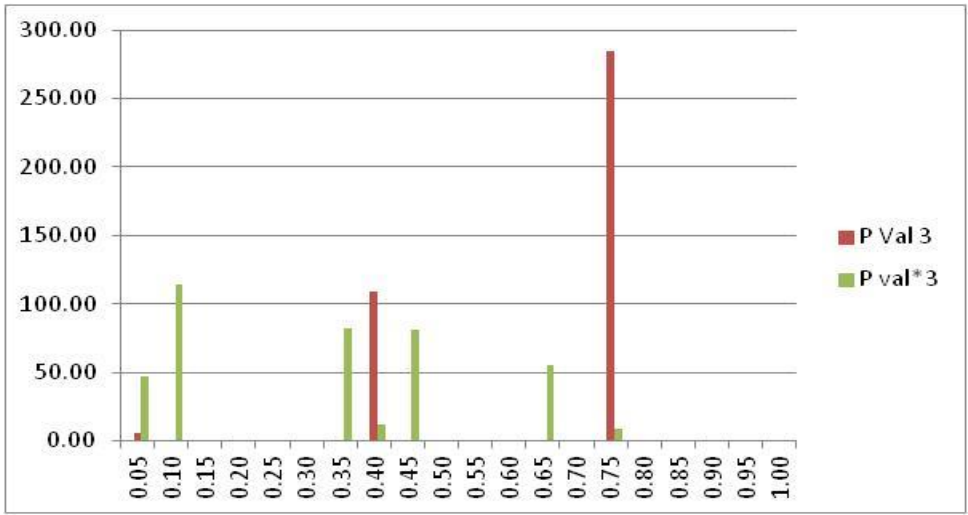
Choose 1st 10 realizations of P and P*; Compute mass functions of P vals for all (252) choices of 5 from the 10



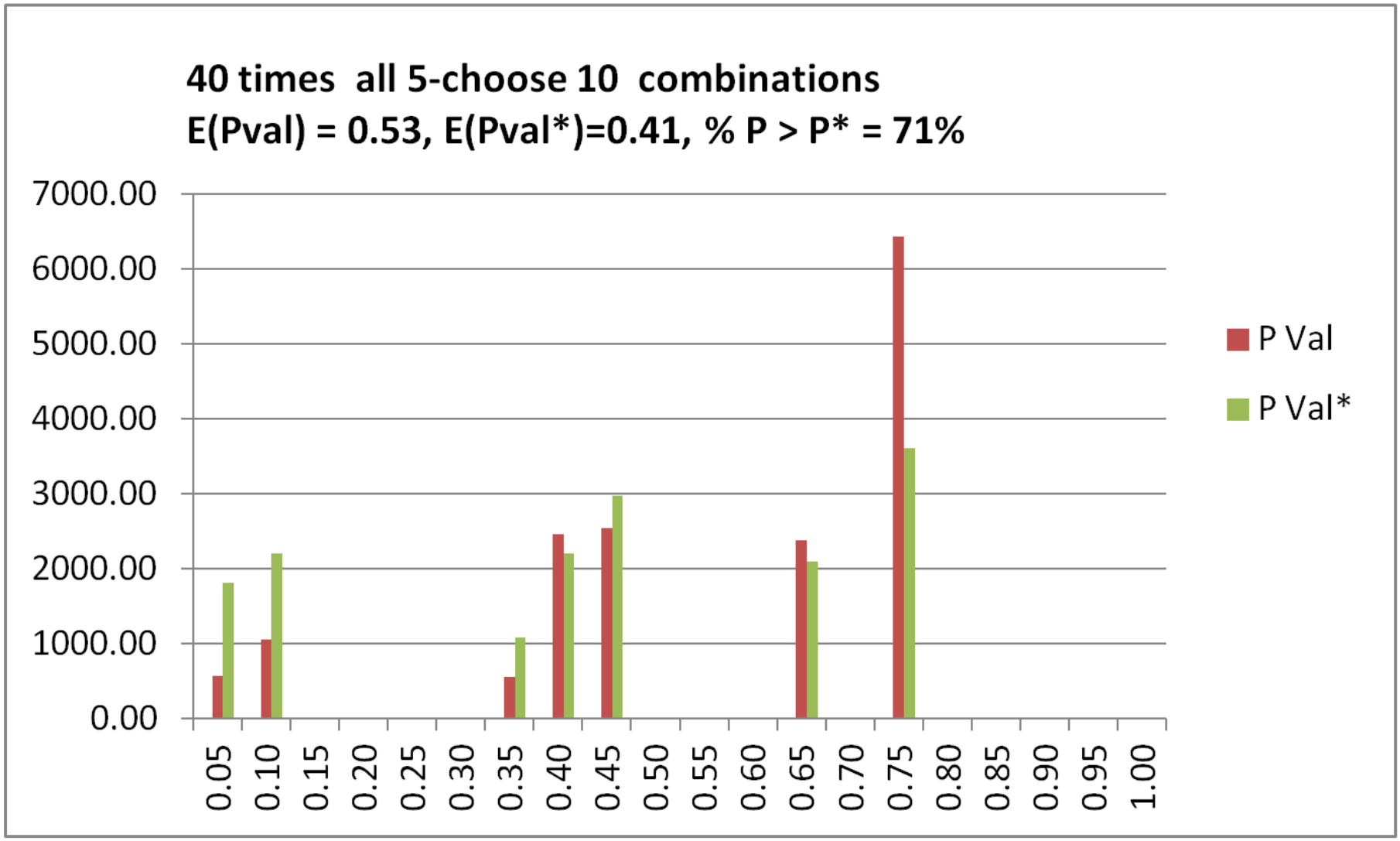
Choose 2nd 10 realizations of P and P*; Compute mass functions of P vals for all choices of 5 from the 10



Choose 3rd 10 realizations of P and P*; Compute mass functions of P vals for all choices of 5 from the 10

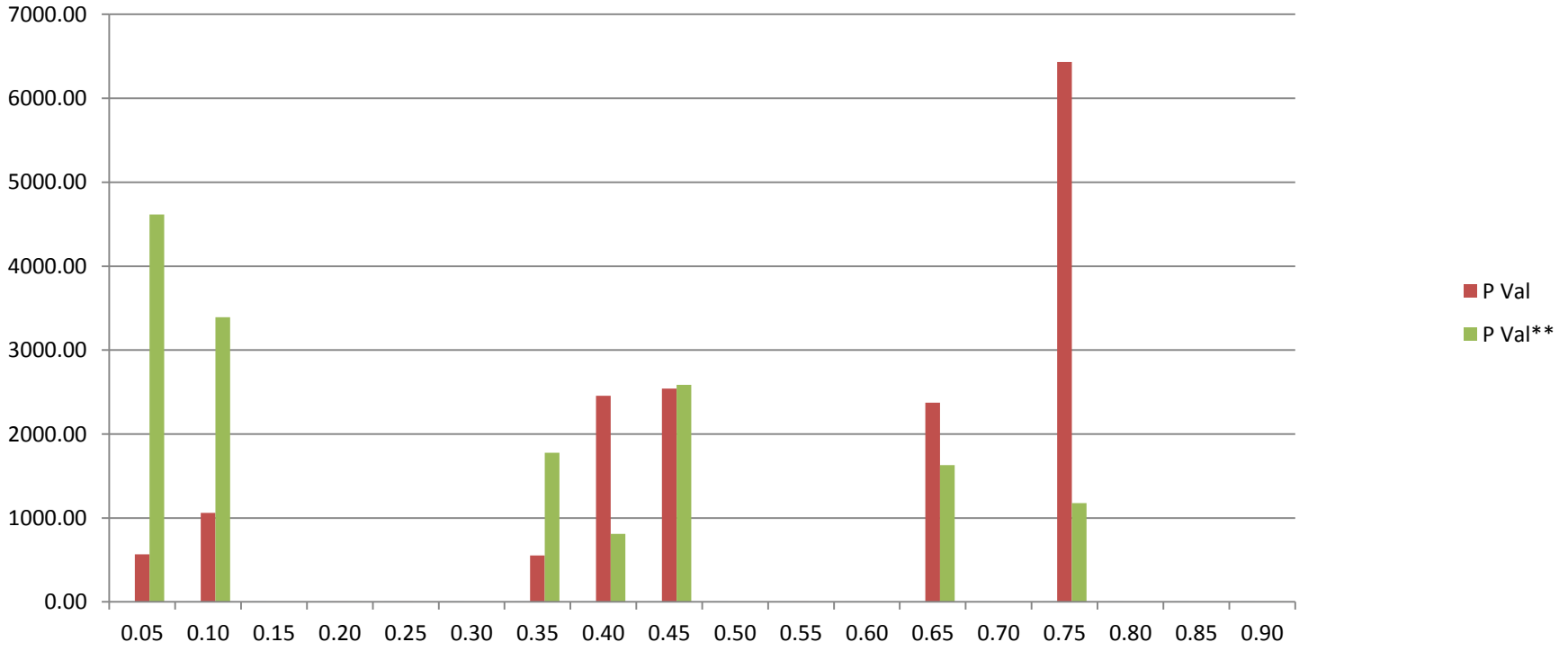


P versus P*

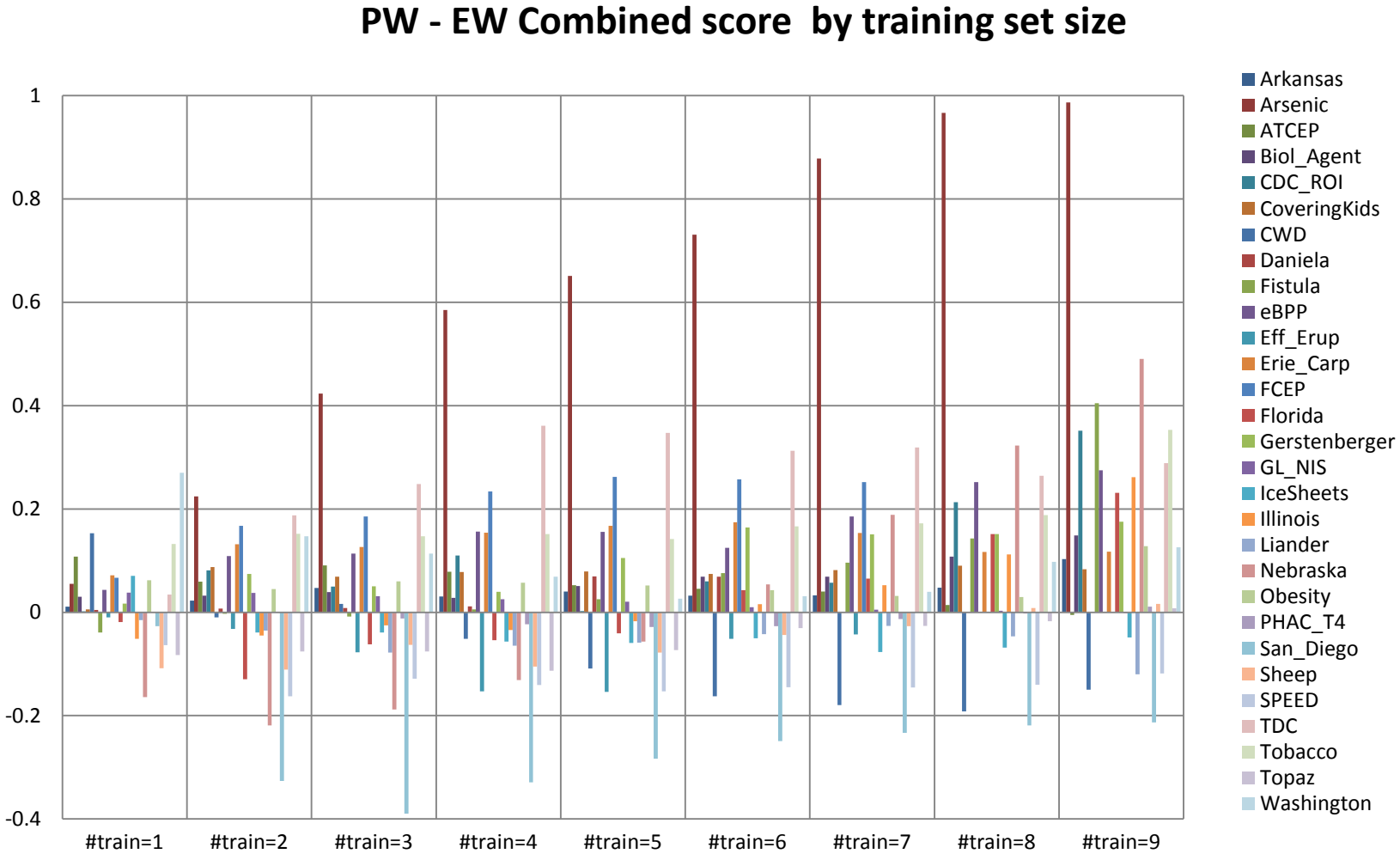

40 times all 5-choose 10 combinations
E(Pval) = 0.53, E(Pval*)=0.41, % P > P* = 71%
7000.00
6000.00
5000.00
4000.00
3000.00
2000.00
1000.00
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
P Val
P Val*

P versus P**

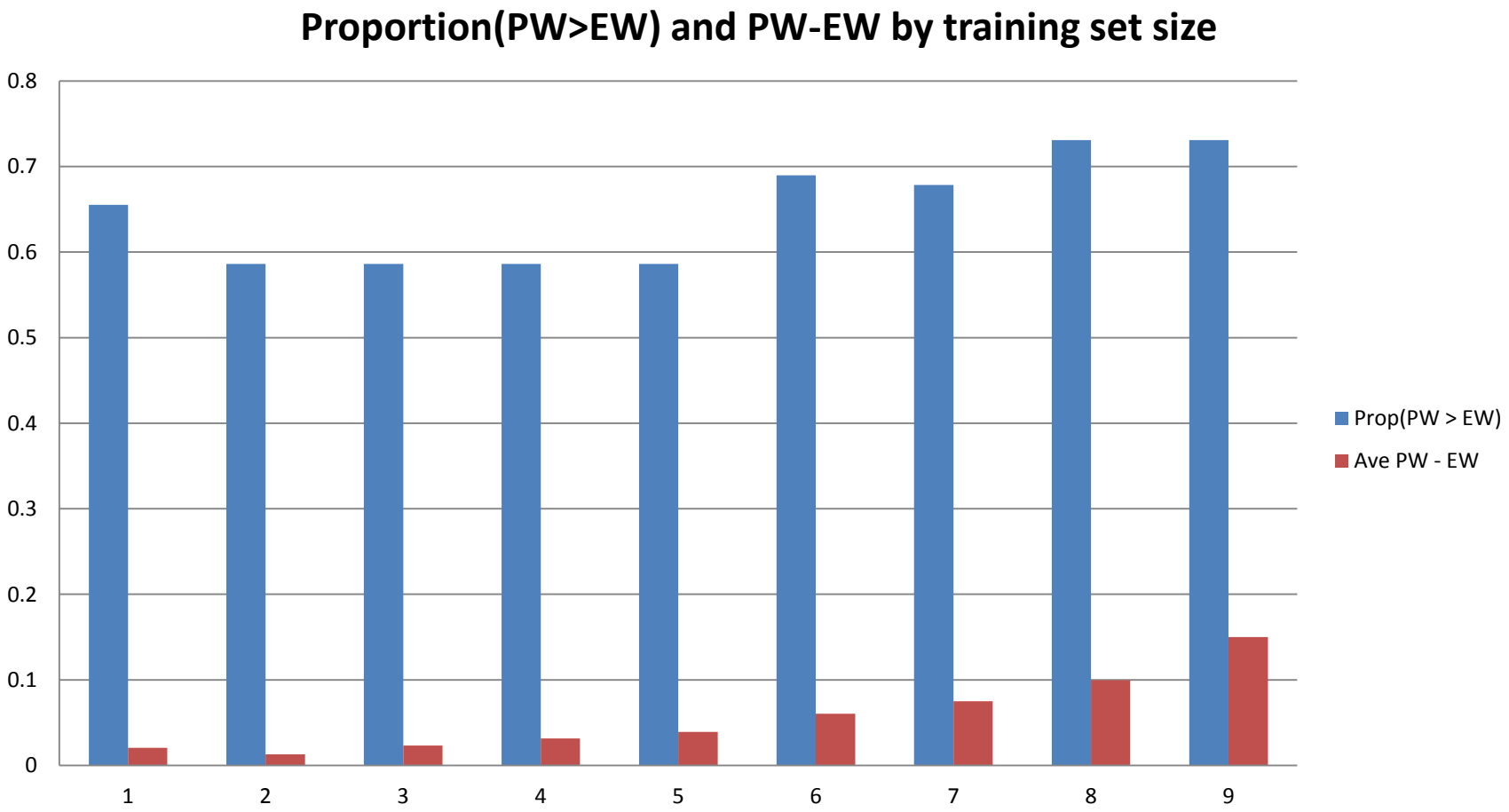
40 times all 5-choose 10 combinations
E(Pval) = 0.53, E(Pval)=0.26), % P > P* = 83%**

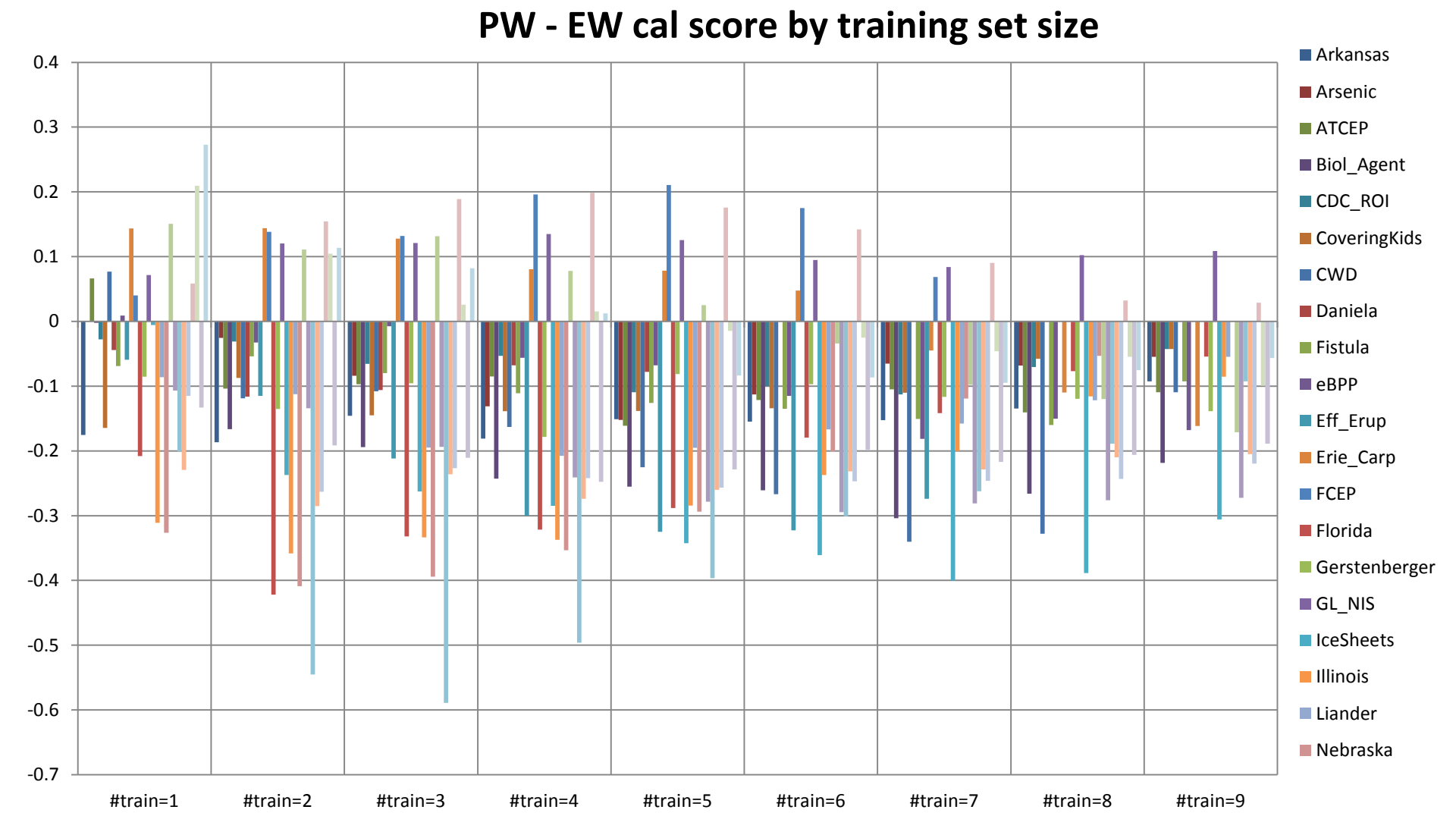


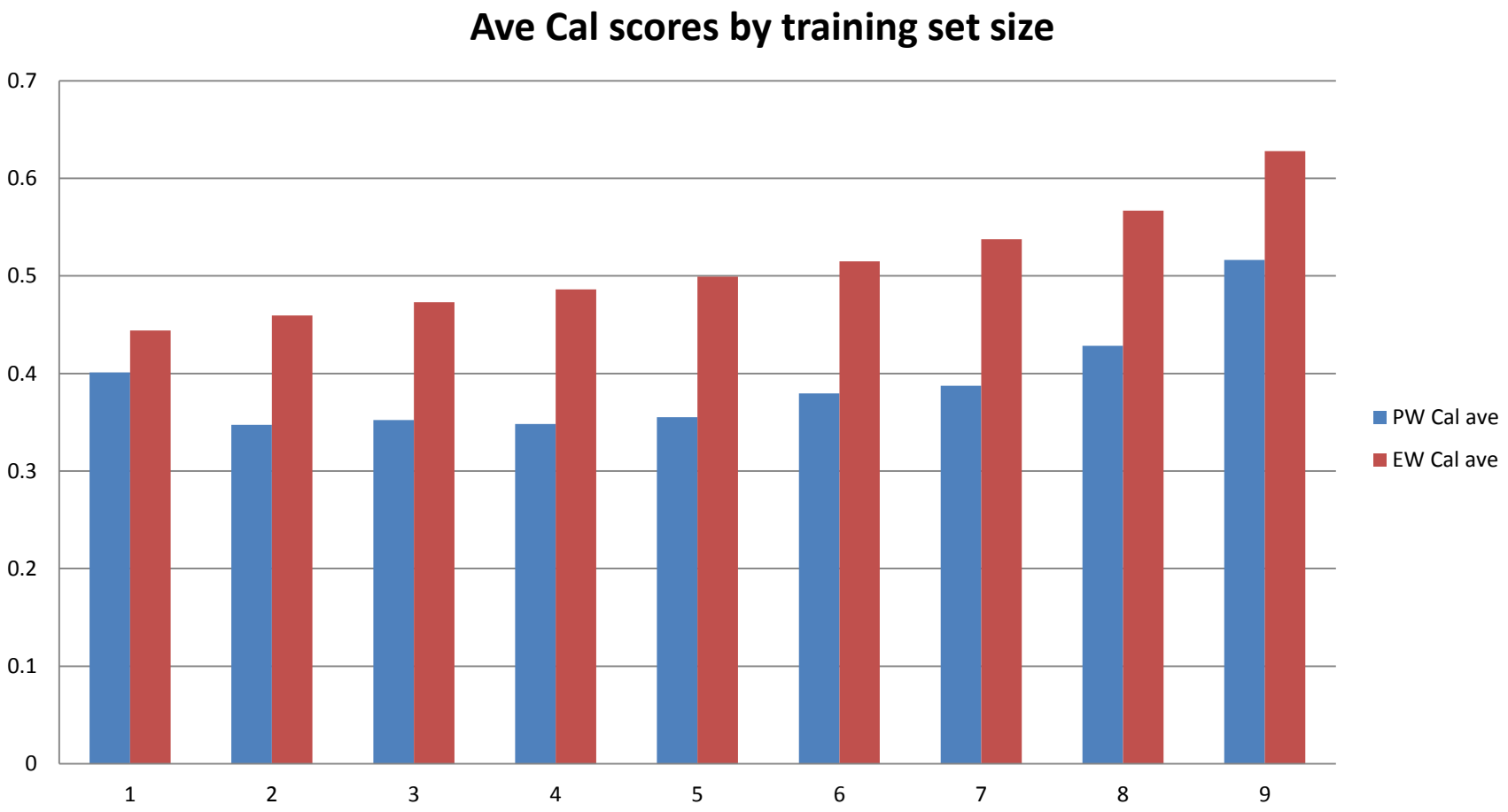
Combined scores study-wise



Over all 29 post 2008 studies

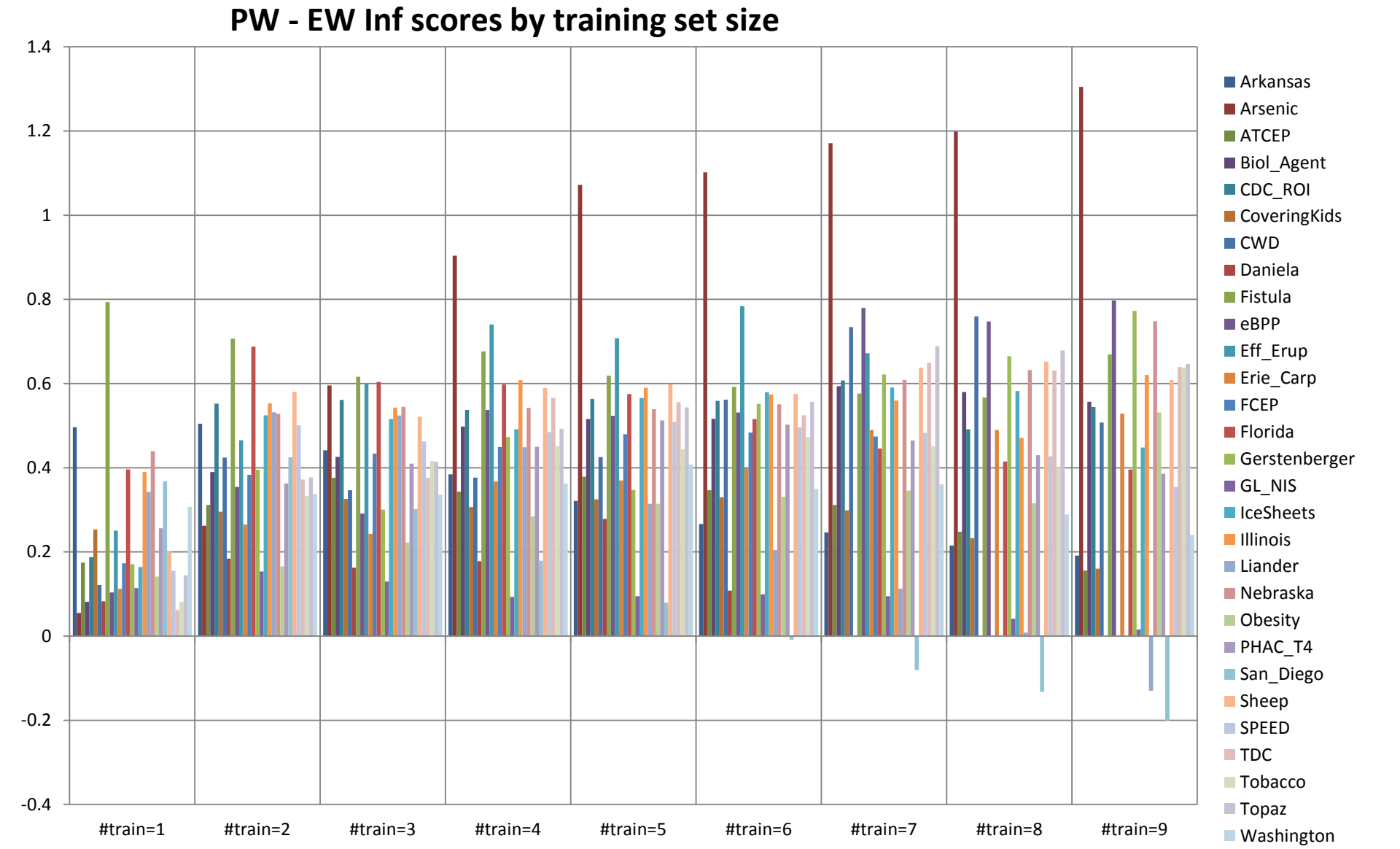



PW - EW cal score by training set size
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
#train=1
#train=2
#train=3
#train=4
#train=5
#train=6
#train=7
#train=8
#train=9
Arkansas
Arsenic
ATCEP
Biol_Agent
CDC_ROI
CoveringKids
CWD
Daniela
Fistula
eBPP
Eff_Erup
Erie_Carp
FCEP
Florida
Gerstenberger
GL_NIS
IceSheets
Illinois
Liander
Nebraska


Ave Cal scores by training set size
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
PW Cal ave
EW Cal ave

In cross validation, a subset and its compliment are anti-correlated

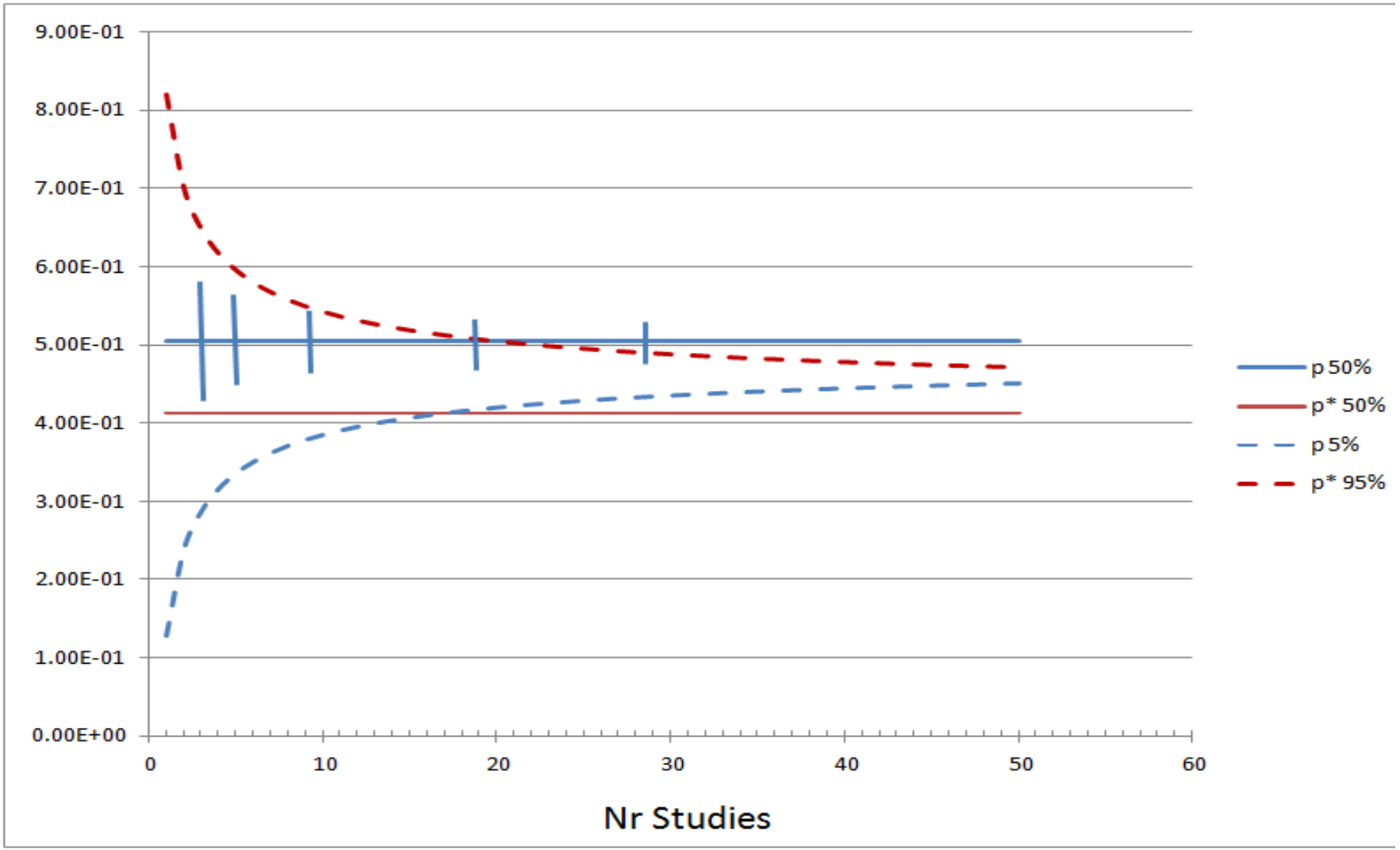
P val correlations of antinimous subsets	
3 out 10	-0.15309
4 out 10	-0.27628
5 out 10	-0.44319
6 out 10	-0.43073
7 out 10	-0.33433
8 out 10	-0.21269


PW - EW Inf scores by training set size
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
#train=1
#train=2
#train=3
#train=4
#train=5
#train=6
#train=7
#train=8
#train=9
Arkansas
Arsenic
ATCEP
Biol_Agent
CDC_ROI
CoveringKids
CWD
Daniela
Fistula
eBPP
Eff_Erup
Erie_Carp
FCEP
Florida
Gerstenberger
GL_NIS
IceSheets
Illinois
Liander
Nebraska
Obesity
PHAC_T4
San_Diego
Sheep
SPEED
TDC
Tobacco
Topaz
Washington

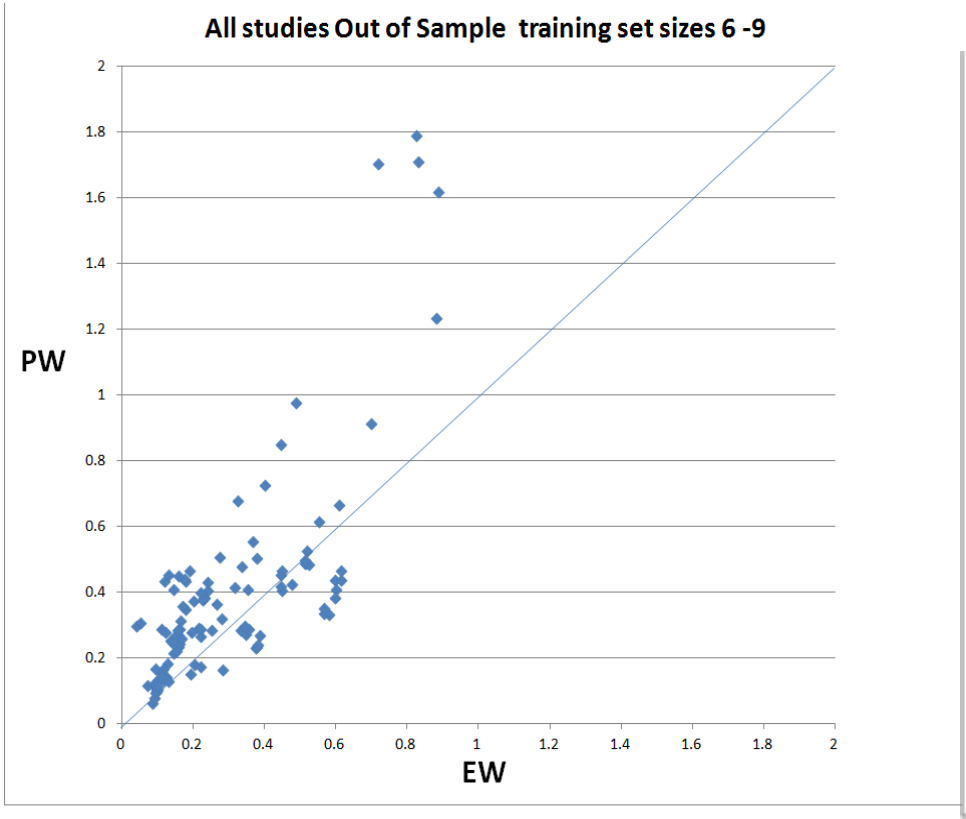

Average Inf scores by training set size
1.2
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
7
8
9
Ave Inf PW
Avw Inf EW

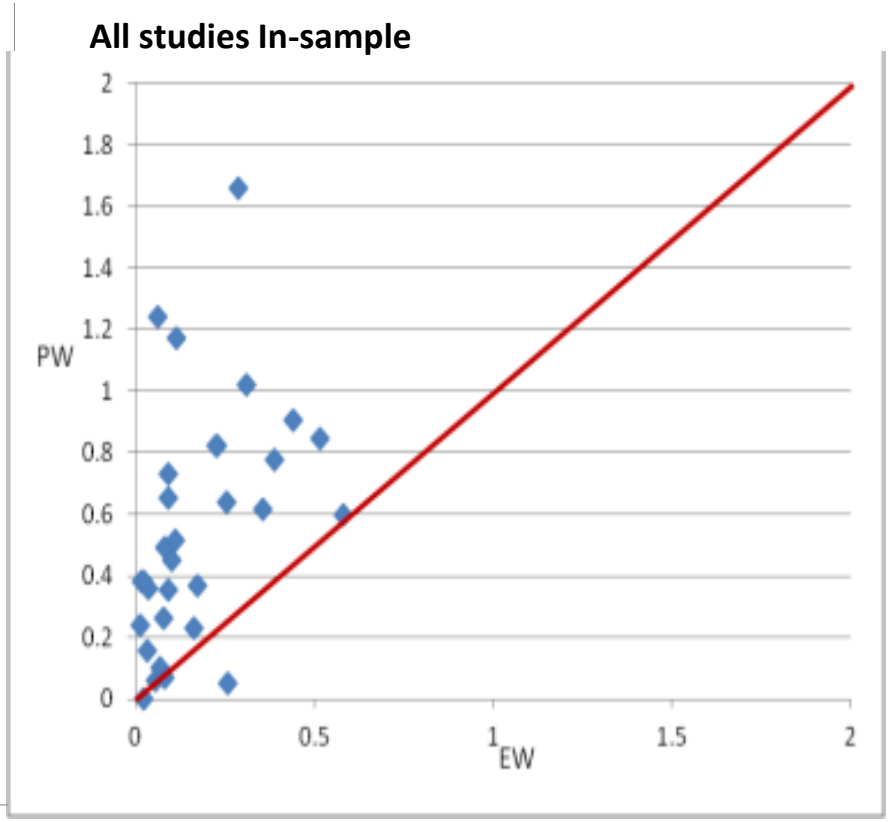
What does cross validation buy?

With 5 out of 10, need ~ 10 study replicates to distinguish P from P*



Summary: out of sample effect:

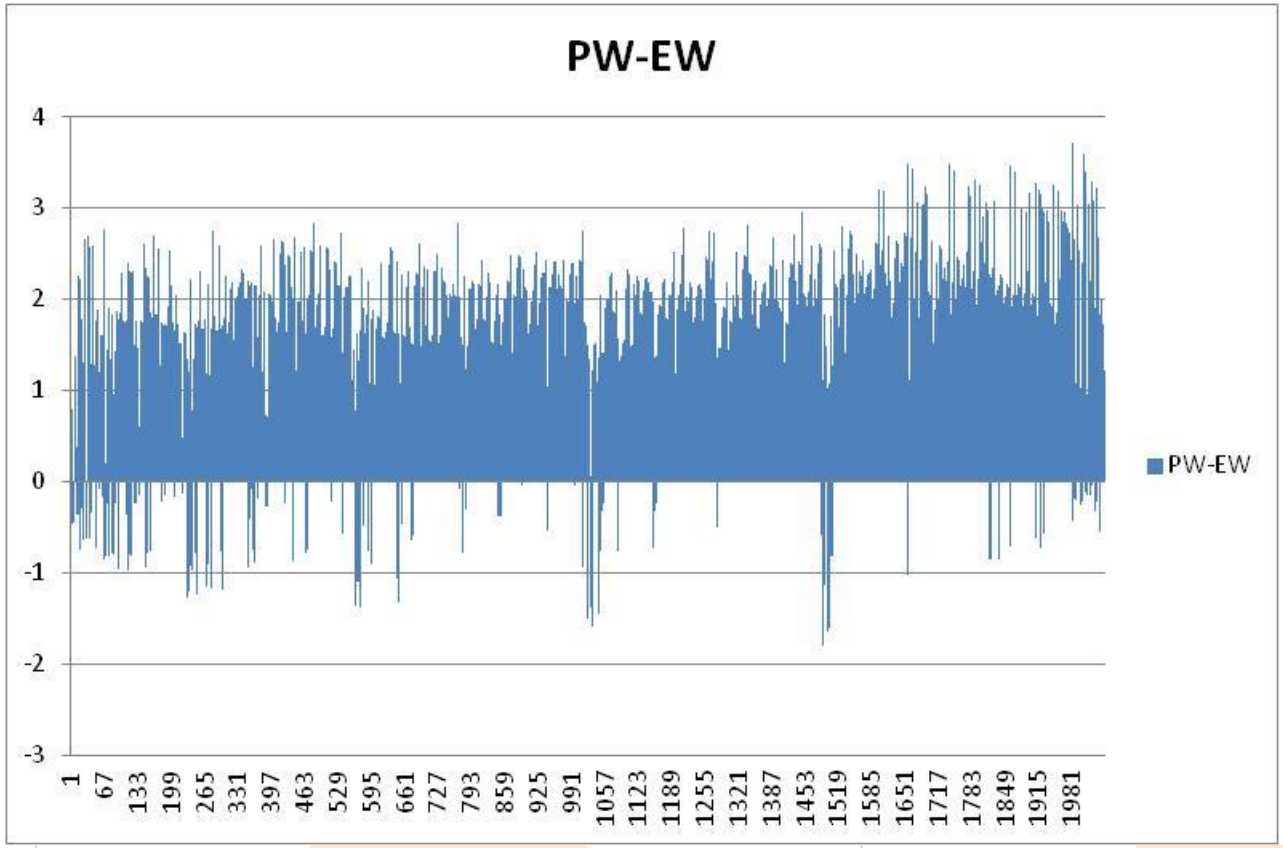




EPA – UMD:

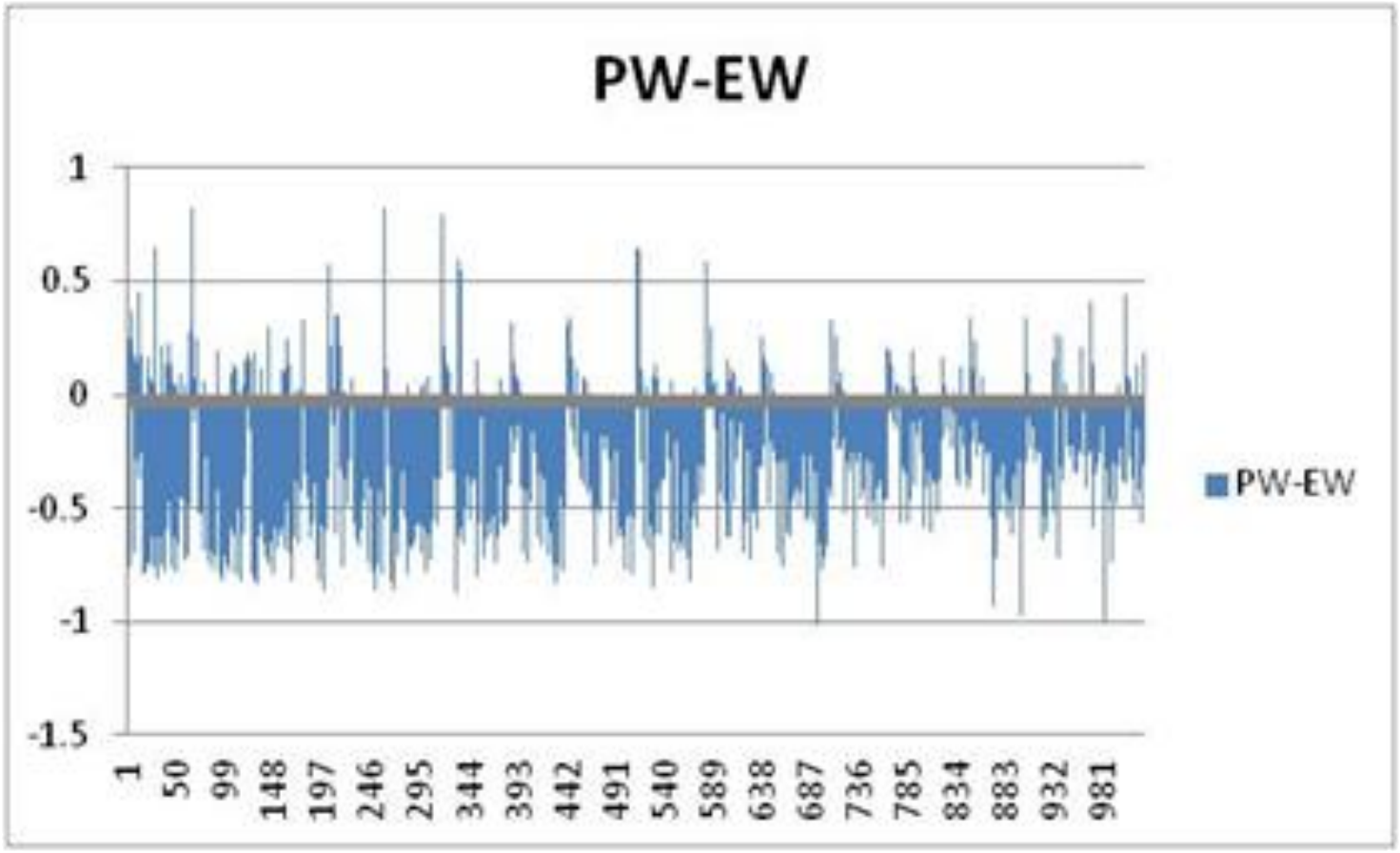
Nitrogen run-off Chesapeake Bay (Palmer, Koch, Febria)

10 experts, 11 seed vbls, 93% PW > EW



EWCal	EWInf	EWComb	PWgCal	PWgInf	PWgComb
0.197	1.809	0.3563	0.7062	4.17	2.945

San Diego RWJF
10 experts, 7 seeds
17%PW>EW



EWCal	EWInf	EWComb	PWgCal	PWgInf	PWgComb
0.3338	1.066	0.356	0.8843	0.6943	0.614

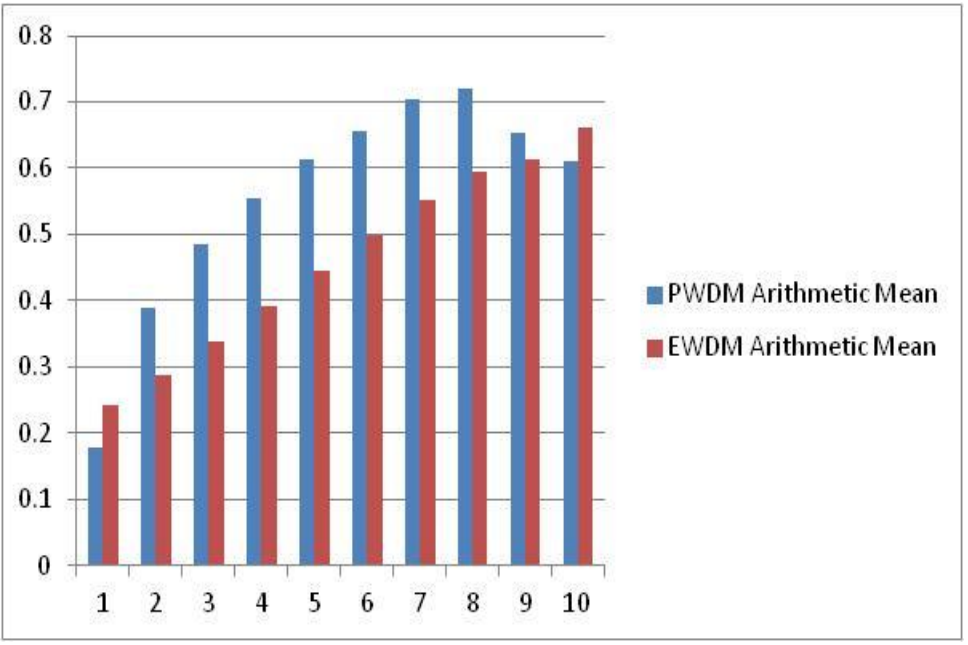
Tentative conclusion

- Default 10 seed vbls is OK to distinguish good and really bad experts, need more for OSV.

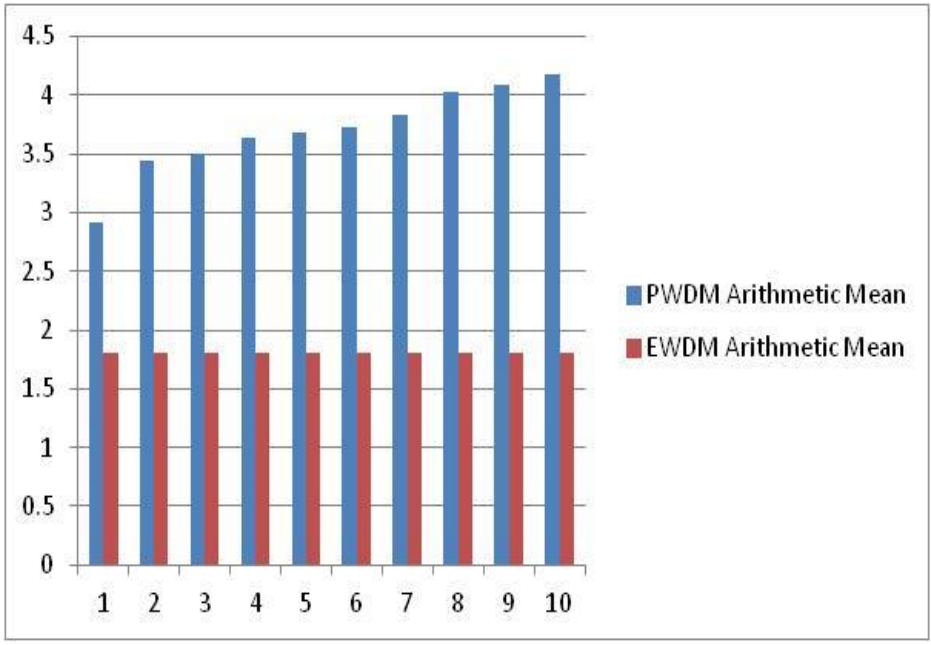
- Large training set is more important than large test set

THANKS for attending

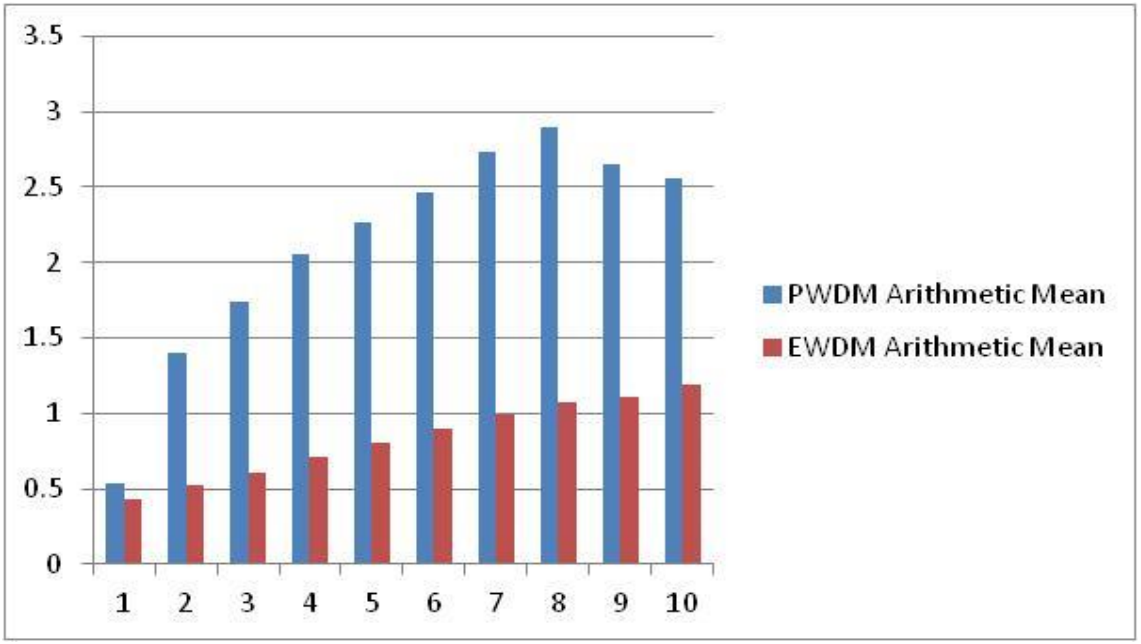
P-value



Information



Combined score



Stdev of P values	
all combo	Stdev
2-out-10	0.10
3-out 10	0.11
4-out-10	0.11
5-out-10	0.11
6-out-10	0.11
7-out-10	0.18
8-out-10	0.14

5-out-10, p	
nr Studies	Stdev
3	0.05438
5	0.04198
10	0.04139
15	0.03062
20	0.02597
30	0.01343
50	0.0035

carp

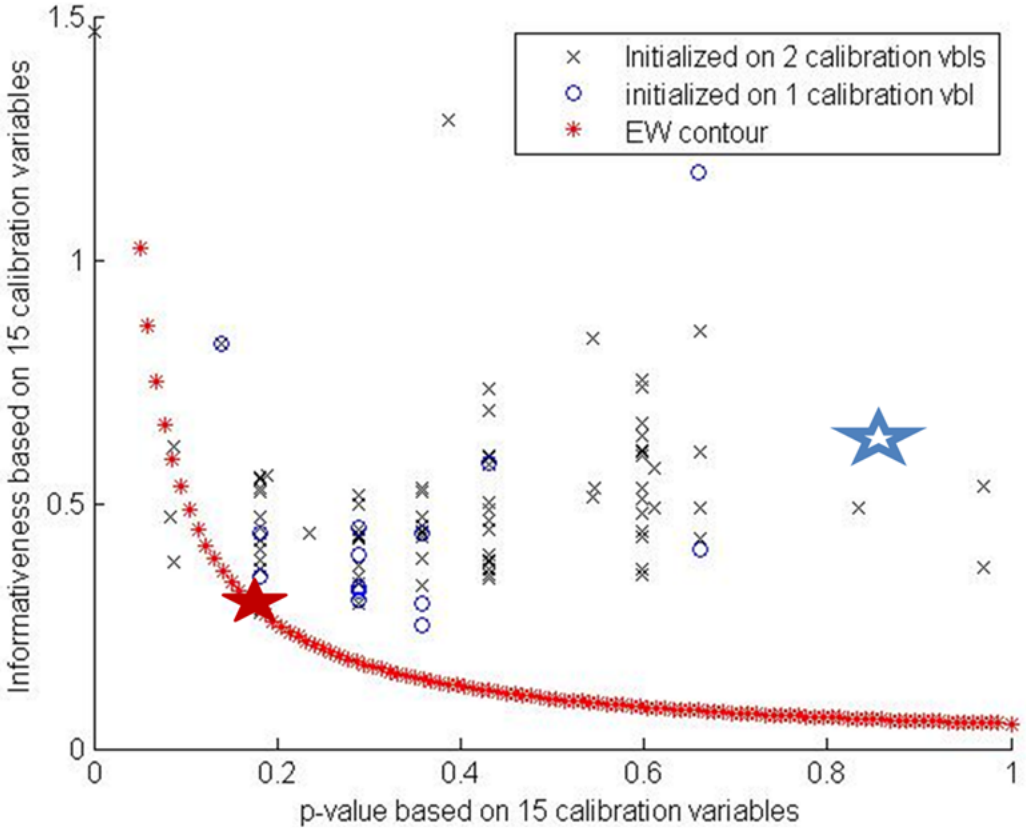


Figure 2: Cross validation results in which the global weight performance based DM is initialized on all subsets of 1 or 2 calibration variables and used to predict all 15 calibration variables. The solid star is the equal weight DM, and the curve consists of combinations of calibration and informativeness resulting in the same combined score as the equal weight DM. The hollow star is the performance DM for all calibration variables.

Ice Sheets

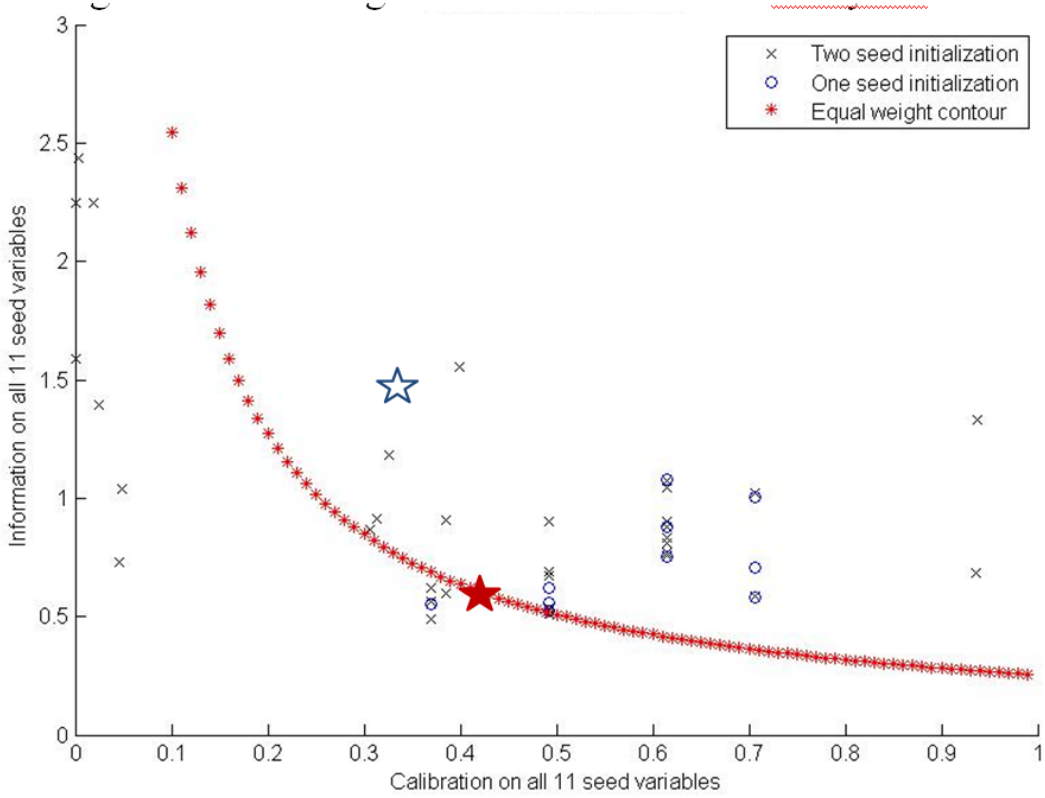


Figure 6 Ice sheet, 11 seed variables: geomean PW/EW one seed = 1.48, two seeds = 0.59. PW > EW on 9 of 11 one seed initializations, and on 34 of 55 two seed initializations, overall on 43/66 = 65% of these initializations.

Obesity

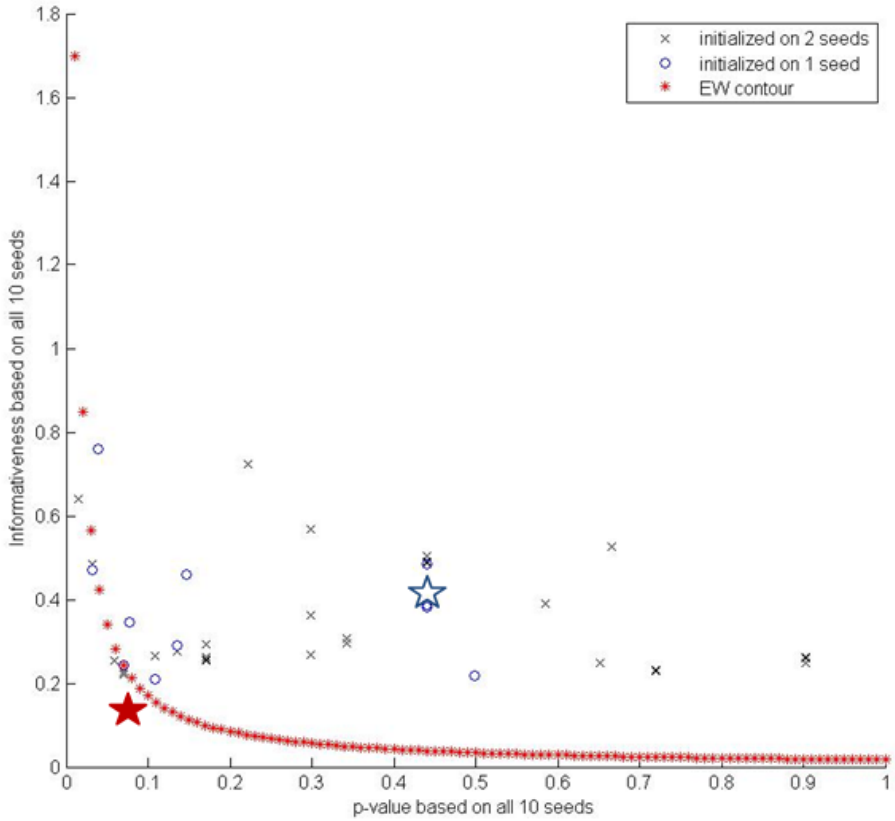


Figure 7: Obesity, 10 seed variables: geomean PW/EW one seed = 3.46, two seeds = 6.19. PW > EW on 9 of 10 one seed initializations, and on 39 of 45 two seed initializations, overall on 48/55 = 87% of these initializations.

Fistula

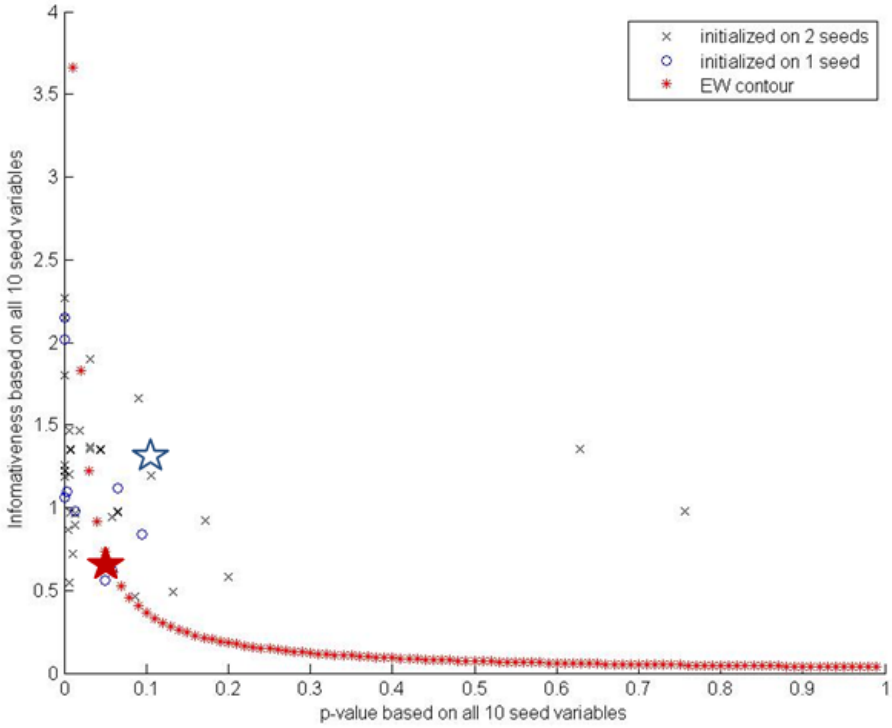


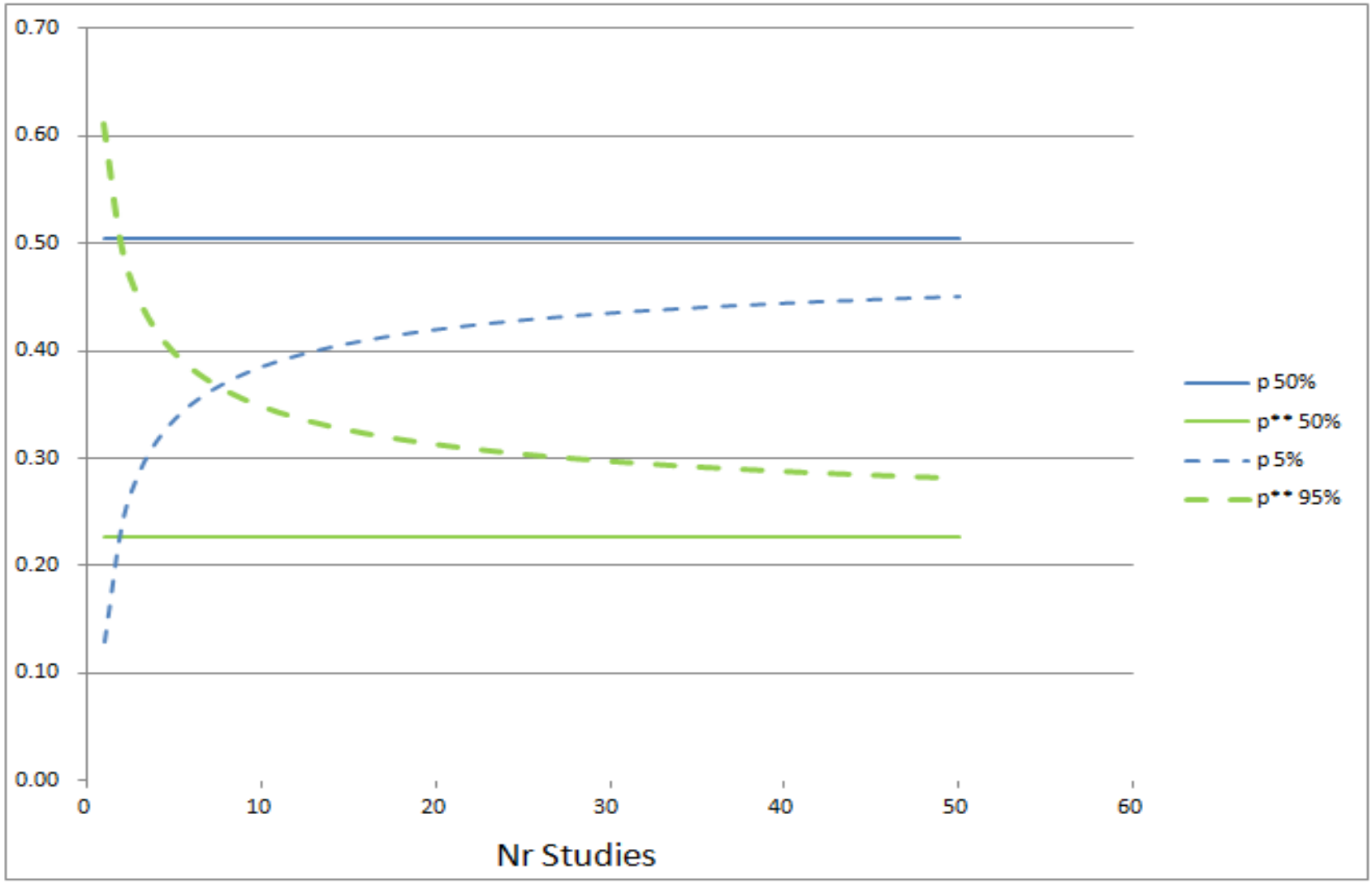
Figure 9: Fistula, 10 seed variables; geomean PW/EW for one seed = 34.8, for two seeds 0.34. PW > EW on 2 of 10 one seed initializations, and on 19 of 45 two seed initializations, overall on 21 / 55 = 38% of these initializations.

DM P is perfectly calibrated

DM P has prob.(30%, 20%, 20%, 30%)** for realizations in [0.05, 0.50, 0.95] interquantile intervals

each study has 5 indep seed vbls.

How many independent studies do we need to distinguish P and P** with 90% confidence?



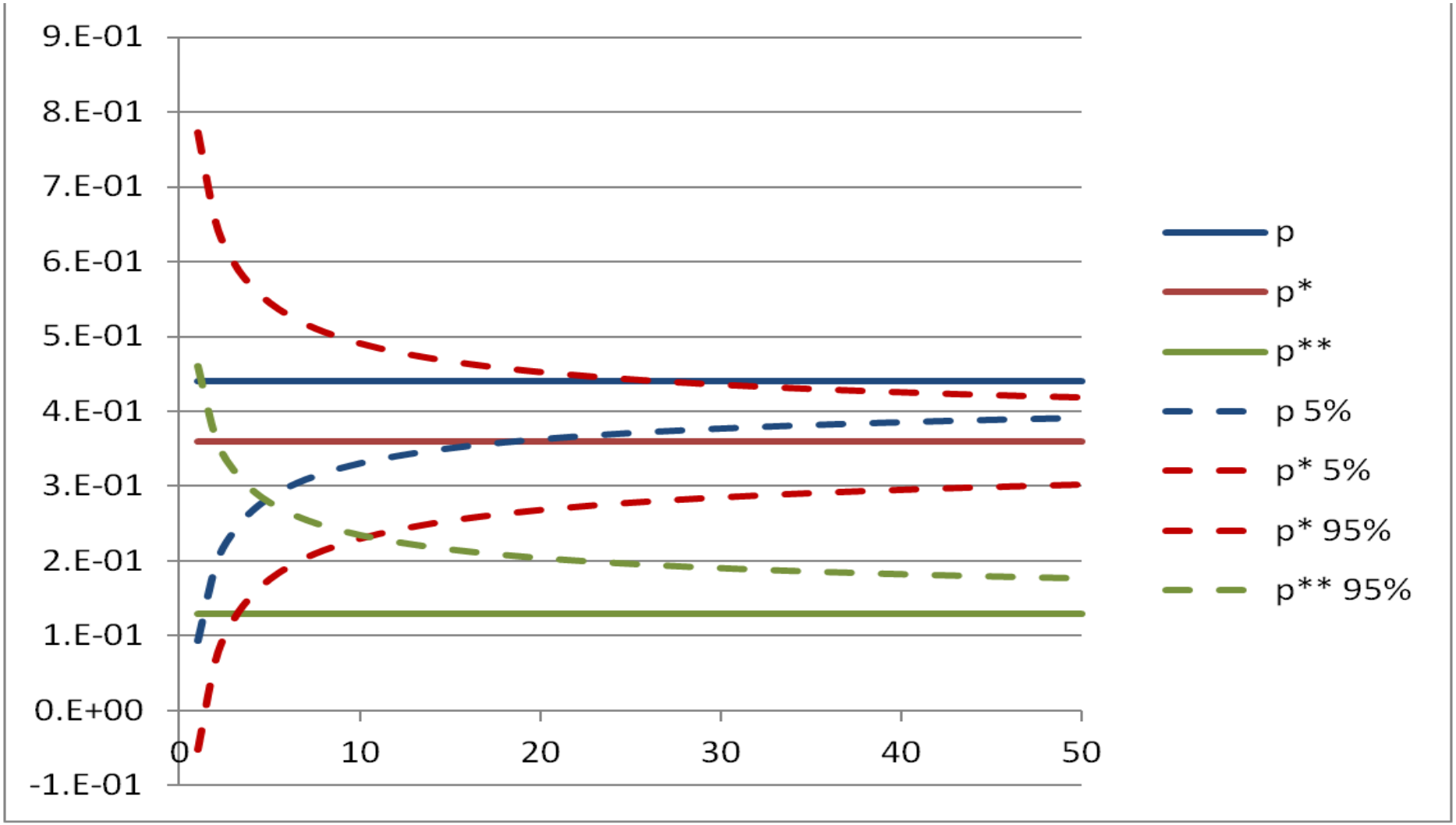
DM P is perfectly calibrated

DM P* has prob.(15%, 30%, 30%, 15%) for realizations in [0.05, 0.50, 0.95] interquantile intervals

DM P has prob.(30%, 20%, 20%, 30%)** for realizations in [0.05, 0.50, 0.95] interquantile intervals

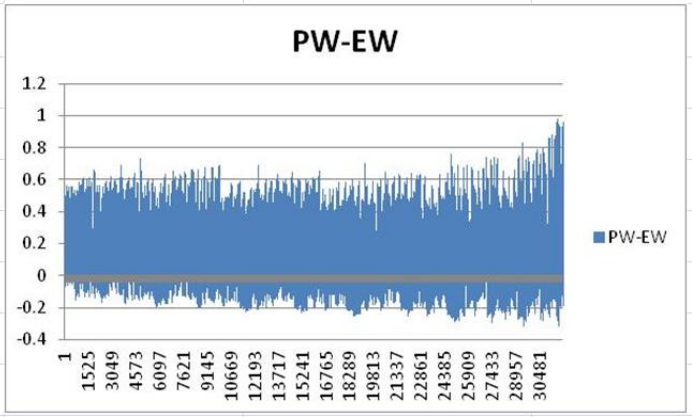
each study has 10 indep seed vbls.

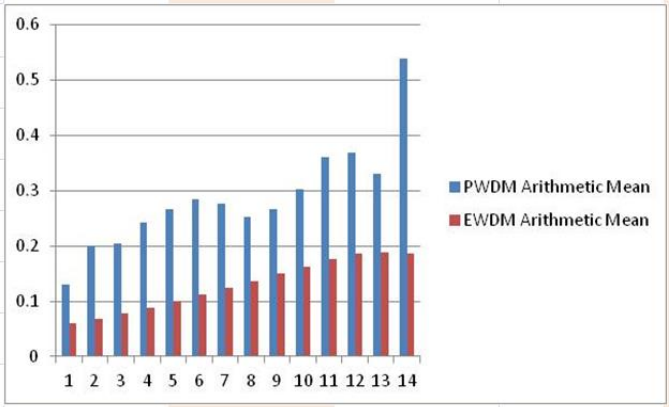
How many independent studies do we need to distinguish P, P* and P** with 90% confidence?



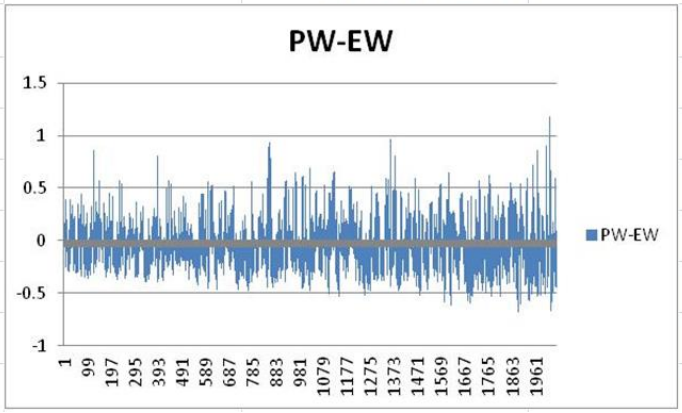
All training sets

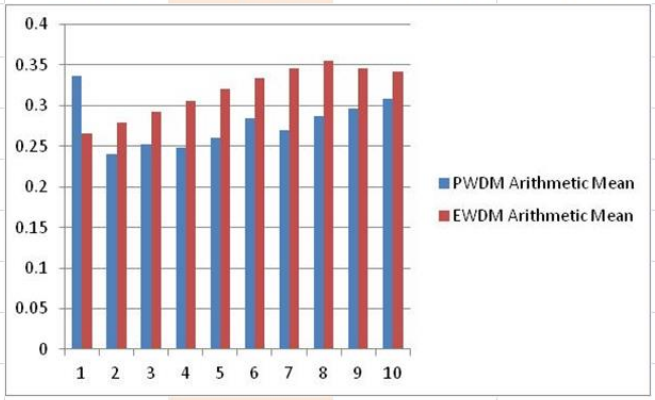
Study	%PW-EW	AVE(PW-EW)	NoOptCal	NoOptInf	NoOptCon	EWCal	EWInf
Erie_Carp	0.74733	0.145032356	0.5683	0.4456	0.2532	0.3126	0.2943
#seeds							
15							
#experts							
10							
by							
TUD							
#quantiles							
3							





Study	%PW-EW	AVE(PW-EW)	NoOptCal	NoOptInf	NoOptCon	EWCal	EWInf
IceSheets	0.37634	-0.057037599	0.615	0.7007	0.4309	0.492	0.5169
#seeds							
11							
#experts							
10							
by							
WPA							
#quantiles							
3							





DO NOT standard scorings rules:

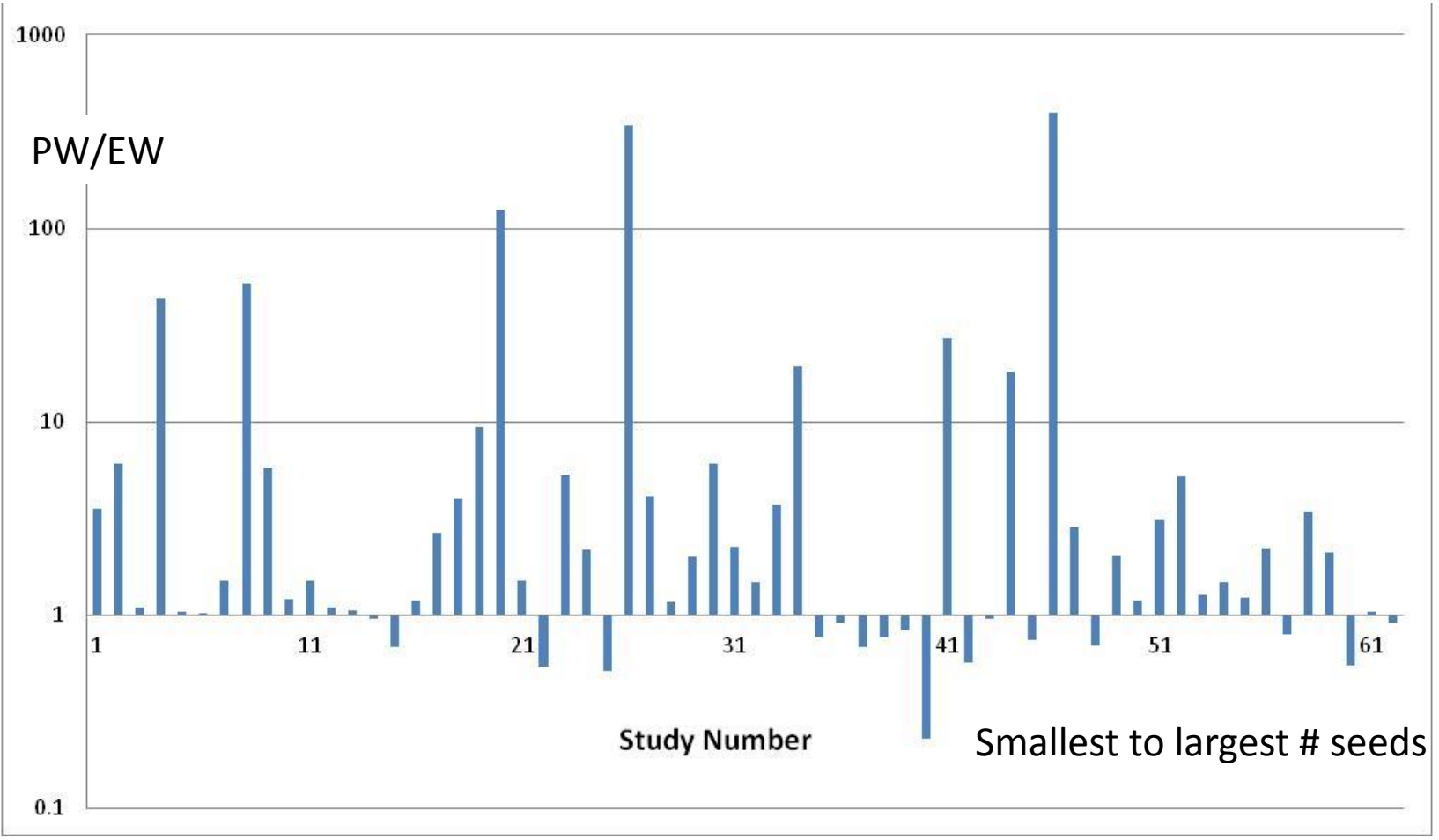
Table 1: Two experts assessing next day probability of rain on 1000 days

Probability of Rain next day:		5%	15%	25%	35%	45%	55%	65%	75%	85%	95%	Totals
expert 1	assessed	100	100	100	100	100	100	100	100	100	100	1000
	realized	5	15	25	35	45	55	65	75	85	95	500
expert 2	assessed	100	100	100	100	100	100	100	100	100	100	1000
	realized	0	0	0	0	0	100	100	100	100	100	500

Quadratic score expert 1 = 0.665; Quadratic score expert 2 = 0.835

Eggstaff et al 2013

PW / EW study-wise



For each K, average PW and EW, take Geomean PW/EW over **all** K= 1 to # seeds.

each study has 5 indep seed vbls.

How many independent studies do we need to distinguish P and P* with 90% confidence?